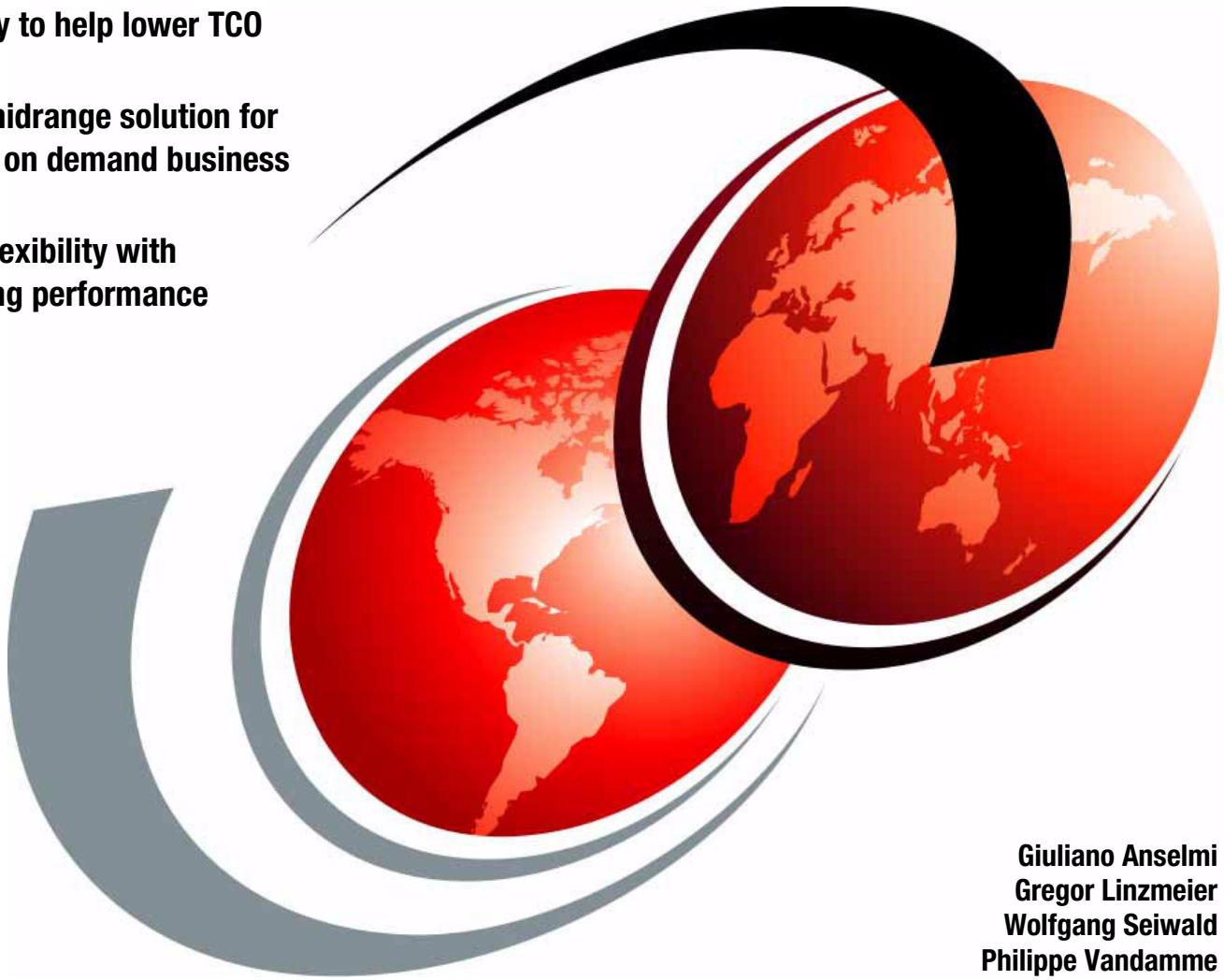
IBM

IBM @server p5 570 Technical Overview and Introduction

- Finer system granulation using Micro-Partitioning technology to help lower TCO
- Modular midrange solution for managing on demand business
- Extreme flexibility with outstanding performance



Giuliano Anselmi
Gregor Linzmeier
Wolfgang Seiwald
Philippe Vandamme

Redpaper

ibm.com/redbooks

International Technical Support Organization

IBM @server p5 570 Technical Overview and Introduction

July 2004

Note: Before using this information and the product it supports, read the information in “Notices” on page vii.

First Edition (July 2004)

This edition applies to the IBM @server p5 570 and AIX 5L™ Version 5.3, product number 5765-G03.

Contents

Notices

This information was developed for products and services offered in the U.S.A.

IBM may not offer the products, services, or features discussed in this document in other countries. Consult your local IBM representative for information on the products and services currently available in your area. Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

IBM may have patents or pending patent applications covering subject matter described in this document. The furnishing of this document does not give you any license to these patents. You can send license inquiries, in writing, to:
IBM Director of Licensing, IBM Corporation, North Castle Drive Armonk, NY 10504-1785 U.S.A.

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law: INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION "AS IS" WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY OR FITNESS FOR A PARTICULAR PURPOSE. Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore, this statement may not apply to you.

This information could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; these changes will be incorporated in new editions of the publication. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this publication at any time without notice.

Any references in this information to non-IBM Web sites are provided for convenience only and do not in any manner serve as an endorsement of those Web sites. The materials at those Web sites are not part of the materials for this IBM product and use of those Web sites is at your own risk.

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

Any performance data contained herein was determined in a controlled environment. Therefore, the results obtained in other operating environments may vary significantly. Some measurements may have been made on development-level systems and there is no guarantee that these measurements will be the same on generally available systems. Furthermore, some measurement may have been estimated through extrapolation. Actual results may vary. Users of this document should verify the applicable data for their specific environment.

Information concerning non-IBM products was obtained from the suppliers of those products, their published announcements or other publicly available sources. IBM has not tested those products and cannot confirm the accuracy of performance, compatibility or any other claims related to non-IBM products. Questions on the capabilities of non-IBM products should be addressed to the suppliers of those products.

This information contains examples of data and reports used in daily business operations. To illustrate them as completely as possible, the examples include the names of individuals, companies, brands, and products. All of these names are fictitious and any similarity to the names and addresses used by an actual business enterprise is entirely coincidental.

COPYRIGHT LICENSE:
This information contains sample application programs in source language, which illustrates programming techniques on various operating platforms. You may copy, modify, and distribute these sample programs in any form without payment to IBM, for the purposes of developing, using, marketing or distributing application programs conforming to the application programming interface for the operating platform for which the sample programs are written. These examples have not been thoroughly tested under all conditions. IBM, therefore, cannot guarantee or imply reliability, serviceability, or function of these programs. You may copy, modify, and distribute these sample programs in any form without payment to IBM for the purposes of developing, using, marketing, or distributing application programs conforming to IBM's application programming interfaces.

Trademarks

The following terms are trademarks of the International Business Machines Corporation in the United States, other countries, or both:

Advanced Micro-Partitioning™
AIX®
AIX 5L™
Chipkill™
Electronic Service Agent™
Enterprise Storage Server®
@server®
@server®
HACMP™
i5/OS™
IBM®
Micro-Partitioning™
POWER™
POWER4™
POWER4+™
POWER5™
PowerPC®
pSeries®
Redbooks™
Redbooks (logo) ™
RS/6000®
Service Director™
TotalStorage®

The following terms are trademarks of other companies:

UNIX is a registered trademark of The Open Group in the United States and other countries.

Linux is a trademark of Linus Torvalds in the United States, other countries, or both.

Other company, product, and service names may be trademarks or service marks of others.

Preface

This document is a comprehensive guide covering the IBM® @server™ p5 570 UNIX® servers. It introduces major hardware offerings and discusses their prominent functions.

Professionals wishing to acquire a better understanding of IBM @server p5 products should read this document. The intended audience includes:

- Customers
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This document expands the current set of IBM @server documentation by providing a desktop reference that offers a detailed technical description of the p5-570 system.

This publication does not replace the latest pSeries® marketing materials and tools. It is intended as an additional source of information that, together with existing sources, may be used to enhance your knowledge of IBM server solutions.

The team that wrote this Redpaper

This Redpaper was produced by a team of specialists from around the world working at the International Technical Support Organization, Austin Center.

Giuliano Anselmi is a certified pSeries Presales Technical Support Specialist working in the Field Technical Sales Support group based in Rome, Italy. For seven years, he was an IBM @server pSeries Systems Product Engineer, supporting Web Server Sales Organization in EMEA, IBM Sales, IBM Business Partners, Technical Support Organizations, and IBM Dublin eServer Manufacturing. Giuliano has worked for IBM for 12 years, devoting himself to RS/6000® and pSeries systems with his in-depth knowledge of the related hardware and solutions.

Gregor Linzmeier is an IBM Advisory IT Specialist for RS/6000 and pSeries workstation and entry servers as part of the Systems and Technology Group in Mainz, Germany supporting IBM sales, Business Partners, and customers with pre-sales consultation and implementation of client/server environments. He has worked for more than 13 years as an infrastructure specialist for RT, RS/6000, and AIX® in large CATIA client/server projects.

Wolfgang Seiwald is an IBM Presales Technical Support Specialist working for the System Sales Organization in Salzburg, Austria. He holds a Diplomingenieur degree in Telematik from the Technical University of Graz. The main focus of his work for IBM in the past five years has been in the areas of the IBM @server pSeries systems and the IBM AIX operating system.

Philippe Vandamme is an IT Specialist working in pSeries Field Technical Support in Paris, France, EMEA West region. With 15 years of experience in semi-conductor fabrication and manufacturing and associated technologies, he is now in charge of pSeries Pre-Sales Support. In his daily role, he supports and delivers training to the IBM and Business Partner Sales force.

The project that produced this document was managed by:

Scott Vetter
IBM U.S.

Thanks to the following people for their contributions to this project:

Ron Arroyo, John Banchy, Barb Hewitt, Thoi Nguyen, Jan Palmer, Charlie Reeves, Craig Shempert, Scott Smylie, Joel Tendler, Ed Toutant, Jane Arbeitman, Tenley Jackson, Andy McLaughlin.
IBM U.S.

Derrick Daines, Dave Williams.
IBM UK

Volker Haug
IBM Germany

Become a published author

Join us for a two- to six-week residency program! Help write an IBM Redbook dealing with specific products or solutions while getting hands-on experience with leading-edge technologies. You'll team with IBM technical professionals, Business Partners, and/or customers.

Your efforts will help increase product acceptance and customer satisfaction. As a bonus, you will develop a network of contacts in IBM development labs, and increase your productivity and marketability.

Find out more about the residency program, browse the residency index, and apply online at:

`ibm.com/redbooks/residencies.html`

Comments welcome

Your comments are important to us. We want our papers to be as helpful as possible. Send us your comments about this Redpaper or other Redbooks™ in one of the following ways:

- Use the online **Contact us** review redbook form found at:

  `ibm.com/redbooks`

- Send your comments in an e-mail to:

  `redbook@us.ibm.com`

- Mail your comments to:

  IBM Corporation, International Technical Support Organization
  Dept. JN9B Building 905 Internal Zip 9053D004
  11501 Burnet Road
  Austin, Texas 78758-3493

1

General description

The IBM @server p5 570 rack-mount server is designed for greater application flexibility, with innovative technology, to capitalize on the e-business revolution at the midrange level for server environments. Introduced with the POWER4™ and POWER4+™ technology in 2001 and available from the 1-way entry-level through the 32-way high-end pSeries systems, the IBM POWER™ architecture achieved a new stage of capability characteristics by including features such as logical partitioning (LPAR). With POWER5™ microprocessor technology, the p5-570 is the first cost-effective, high-performance midrange UNIX server to include the next development of the IBM partitioning concept, Micro-Partitioning™.

Dynamic logical partitioning is supported from the 2-way p5-570 to the 16-way p5-570 system, allowing up to 16 dedicated partitions. In addition, the optional Advanced POWER Virtualization hardware feature enables a technology called Micro-Partitioning technology. The p5-570 system has been designed to support up to 160 partitions on a 16-way system. The Micro-Partitioning technology is an advanced feature of the POWER5 processor that enables multiple partitions to share a physical processor. The extended POWER Hypervisor controls dispatching the physical processors to each of the partitions using Micro-Partitioning technology. In addition to Micro-Partitioning technology, the Advanced POWER Virtualization feature enables sharing of network and storage adapters to satisfy the I/O requests of partitions that do not have a dedicated physical I/O adapter.

In combination with the extraordinary POWER5 processor, the Micro-Partitioning technology is designed to increase system management efficiency and lowers operating expenses through the multiple use of single physical resources that are installed in the p5-570 system.

Simultaneous multi-threading (SMT), a standard feature of POWER5 technology, enables two threads to be executed at the same time on a single processor. SMT is user-selectable with dedicated or processors from a shared pool for use by partitions using Micro-Partitioning technology.

The symmetric multiprocessor (SMP) p5-570 system features base 2-way, 4-way, 8-way, 12-way, and 16-way, 64-bit, copper-based, SOI-based POWER5 microprocessors running at 1.5 GHz, 1.65 GHz, and 1.9 GHz with 36 MB off-chip Level 3 cache configurations. The system is based on a concept of system building blocks. The p5-570 building blocks are facilitated with the use of Processor interconnect and system SP Flex cables that enable as many as four 4-way p5-570 building blocks to be connected to achieve a true 16-way SMP

combined system. Additional processor configurations are possible with the installation of Capacity on Demand (CoD) features. Main memory starting at 2 GB can be expanded to 128 GB in a single drawer, based on the available DIMMs, for higher performance and exploitation of 64-bit addressing, to meet the demands of enterprise computing, such as large database applications.

One p5-570 building block includes six hot-plug PCI-X[1] slots with Enhanced Error Handling (EEH) and an enhanced blind-swap mechanism, two Ultra320 SCSI controllers, one 10/100/1000 Mbps integrated dual-port Ethernet controller, two serial ports, two USB 2.0 ports, two HMC ports, two remote RIO-2 ports, and two System Power Control Network (SPCN) ports.

The p5-570 includes two 3-pack front-accessible, hot-swap-capable disk bays. The six disk bays of one IBM @server p5-570 building block can accommodate up to 880.8 GB of disk storage using the 146.8 GB Ultra320 SCSI disk drives. Two additional media bays are used to accept optional slim-line media devices, such as DVD-ROM or DVD-RAM drives. The p5-570 also has I/O expansion capability using the RIO-2 bus, which allows attachment of the 7311 Model D10, 7311 Model D11, and 7311 Model D20 I/O drawers.

Additional reliability and availability features include redundant hot-plug cooling fans and redundant power supplies. Along with these hot-plug components, the p5-570 is designed to provide an extensive set of reliability, availability, and serviceability (RAS) features that include improved fault isolation, recovery from errors without stopping the system, avoidance of recurring failures, and predictive failure analysis.

[1] PCI stands for Perhiperal Component Interconnect, and the X stands for extended performances.

1.1 System specifications

Table 1-1 lists the general system specifications of a single p5-570 drawer.

Table 1-1 p5-570 specifications

Description	Range
Operating temperature	5 to 35 degrees C (41 to 95 F)
Relative humidity	8% to 80%
Maximum wet bulb	23 degrees C (73 F) (operating)
Noise level	6.5 bels (operating)
Operating voltage	200 to 240 V AC 50/60 Hz
Maximum power consumption	1,300 watts (maximum)
Maximum power source loading	1.37 kVA (maximum configuration)
Maximum thermal output	4,437 Btu[a]/hr (maximum configuration)

a. British Termal Unit (BTU)

1.2 Physical package

One p5-570 drawer is packaged in a 4U[2] rack-mounted enclosure, and it is available only in the rack-mounted form factor. The following sections discuss the major physical attributes that are found on the p5-570 building block, as shown in Table 1-2 on page 3.

Table 1-2 Physical packaging of the p5-570

Dimension	One p5-570 building block
Height	174.1 mm (6.85 in)
Width	483 mm (19.0 in)
Depth	790 mm (31.1 in)
Weight	63.6 kg (140 lb)

Using the p5-570 building block, an installed system can be made of one to four building blocks. To help ensure the installation and serviceability in non-IBM, industry-standard racks, review the vendor's installation planning information for any product-specific installation requirements. The processor and SP Flex cables present an additional planning requirement.

[2] One Electronic Industries Association Unit (1U) is 44.45 mm (1.75 in).

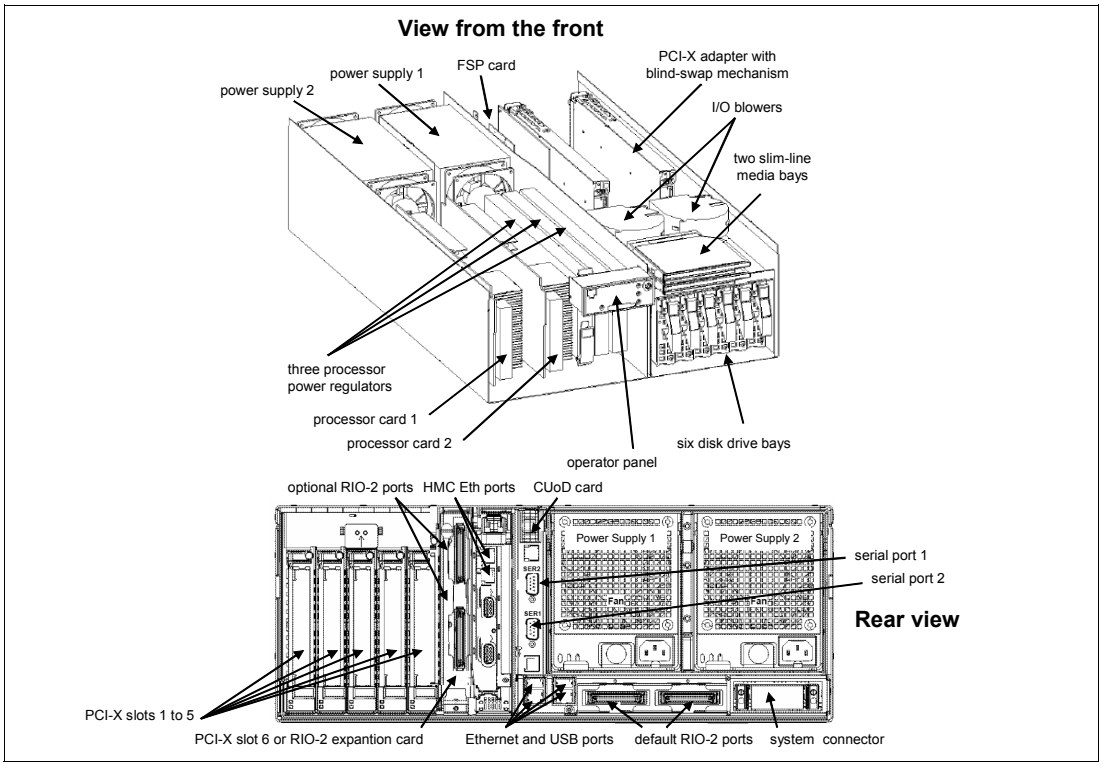


Figure 1-1 Views of the p5-570

1.3 Minimum and optional features

The p5-570 full configuration system is made of four p5-570 building blocks. It features:

- Up to eight processor books using the POWER5 chip, for a total of 16 processors
- From 2 GB to 512 GB of total system memory capacity using DDR1 DIMM technology, or from 2 GB to 64 GB total memory with DDR2 DIMM technology in a four-drawer system
- 24 SCSI disk drives for an internal storage capacity of 3.5 TB using 146.8 GB drives
- 24 PCI-X slots
- Eight slim-line media bays for optional optical storage devices

The combined system (made of more than one p5-570 building block) requires the proper Processor interconnect cable and the system SP Flex cable. (See 2.2.1, “Processor drawer interconnect cables” on page 24, and 2.4.2, “SP bus” on page 28.)

The p5-570 building block includes the service processor (SP), which is described in 2.10, “Service processor” on page 41, and the following native ports:

- Two 10/100/1000 Ethernet ports
- Two serial ports
- Two USB 2.0 ports
  - Optional external USB diskette drive 1.44 (FC 2591) can be used.
- Two HMC ports
- Two remote I/O (RIO-2) ports

- Two SPCN ports

In addition, the p5-570 building block features two internal Ultra320 SCSI controllers, redundant hot-swap power supply and redundant hot-swap cooling fans, and redundant processor power regulators (FC 7875).

There is a CUoD card as part of the hardware configuration. This card stores VPD, and processor information required for management of CUoD features. Since the p5-570 can have processors in up to four physical building blocks, the card can be replaced or updated by the IBM service representative to reflect hardware configuration changes.

Note: In a p5-570 combined system made of more than one building block, only the two HMC ports and the two serial ports in the building block with the service processor are available to use.

The system supports 32-bit and 64-bit applications, and it requires specific levels of operating system. (See 2.12, “Operating system requirements” on page 49.)

1.3.1 Processor card features

Each p5-570 building block can contain 2-way processor cards with state-of-the-art, 64-bit, copper-based, POWER5 microprocessors running at 1.5 GHz, 1.65 GHz, or 1.9 GHz. The processor cards running at 1.5 GHz support up to eight processors per combined system. The 1.65 GHz and 1.9 GHz processor card features are available only as Capacity Upgrade on Demand (CUoD). The initial order of the p5-570 system must contain the feature code related to the desired processor card, plus it must contain the processor activation feature code. The feature can be *no-additional-charge CUoD* (one processor no additional charge, belonging to Value Pak options; see 1.4, “Value Paks” on page 11), *CUoD* (shipped from manufacturing as activated, or for later activation of available non-activated processors), *reserve CoD* activation of prepaid processor, and *On/Off CoD* activation to use the On/Off CoD capabilities. See 3.1, “Capacity on Demand” on page 52 for further details about CoD activation concepts. Table 1-3 and Table 1-4 on page 5 contain all of the available feature codes for processor cards that were available at the time of writing.

Table 1-3 Processor card feature codes

Processor card FC	Description
7834	Two processors, two activated[a], 1.5 GHz, eight DDR1 DIMM sockets
7830	Two processors, 0 activated, 1.65 GHz, eight DDR1 DIMM sockets
7832	Two processors, 0 activated, 1.9 GHz, eight DDR1 DIMM sockets
7833	Two processors, 0 activated, 1.9 GHz, eight DDR2 DIMM sockets

a. The 1.5 GHz processor card does not support CUoD. However, for ordering, any FC 7834 requires two entitlements. The entitlement features are available either full priced or no additional charge (when ordering in a Value Pak).

Table 1-4 Processor activation feature codes

Processor card FC	Description			
One processor activation	**For FC 7834**	**For FC 7830**	**For FC 7832**	**For FC 7833**
No-additional charge CUoD	FC 8456	FC 8452	FC 8454	FC 8455

Processor card FC	Description			
CUoD (permanent)	FC 7929	FC 7897	FC 7898	FC 7899
Reserve CoD	not available	FC 7956	FC 7959	FC 7959
On/Off CoD (1-day billing)	not available	FC 7951, 7952	FC 7951, 7953	FC 7951, 7955

Each processor card features one POWER5 chip, with two processor cores that share 1.9 MB of L2 cache, 36 MB of L3 cache, eight slots for memory DIMMs using DDR1 or DDR2 technology, and requires a minimum of 2 GB memory. (See “Memory features.”)

1.3.2 Memory features

The processor cards that are used in the p5-570 system offer eight sockets for memory DIMMs. The total memory capacity requires four p5-570 building blocks and eight processor cards. DDR1 DIMM and DDR2 DIMM are different technologies that require different memory sockets, so the processor card with POWER5 microprocessors running at 1.9 GHz is available with two feature codes to allow the two different memory technologies. Table 1-5 shows the memory feature codes that are available at the time of writing. The p5-570 system supports CUoD options for memory. (See 3.1.2, “Capacity Upgrade on Demand for memory” on page 53 for more details.)

Table 1-5 Memory feature codes

Feature code	Description
4452	2048 MB (4x512 MB) DIMMs, 208-pin, 8 ns Stacked DDR1 SDRAM
4454	8192 MB (4x2048 MB) DIMMs, 208-pin, 8 ns Stacked DDR1 SDRAM
4490	4096 MB (4x1024 MB) DIMMs, 208-pin, 250 MHz Stacked DDR1 SDRAM
4491	16384 MB (4x4096 MB) DIMMs, 208-pin, 250 MHz Stacked DDR1 SDRAM
4492	32768 MB (4x8192 MB) DIMMs, 208-pin, 250 MHz Stacked DDR1 SDRAM
7892	2048 MB (4x512 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM
7893	4096 MB (4x1024 MB) DIMMs, 276-pin, 533 MHz DDR2 SDRAM

It is recommended that each processor card have an equal amount of memory installed. Balancing memory across the installed processor cards enables memory accesses to be distributed evenly over system components to provide optimal performance.

1.3.3 Disk and media features

Each p5-570 building block features six disk drive bays and two slim-line media device bays. In a full configuration with four connected p5-570 building blocks, the combined system supports up to 24 disk bays; therefore, the maximum internal storage capacity is 3.5 TB (using the disk drive features available at the this time of writing). The minimum configuration requires at least one 36.4 GB disk drive. Table 1-6 shows the disk drive feature codes that each bay can contain.

Table 1-6 Disk drive feature code description

Feature Code	Description
3273	36.4 GB, 10K RPM Ultra320 SCSI disk drive assembly

Feature Code	Description
3277	36.4 GB 15K RPM Ultra320 SCSI disk drive assembly
3274	73.4 GB 10K RPM Ultra320 SCSI disk drive assembly
3278	73.4 GB 15K RPM Ultra320 SCSI disk drive assembly
3275	146.8 GB 10K RPM Ultra320 SCSI disk drive assembly

In a full configuration, with four connected p5-570 building blocks, the combined system supports up to eight slim-line media device bays. To support two slim-line devices in each p5-570 building block, the optional media enclosure and backplane (FC 7869) is required.

Any combination of the following DVD-ROM and DVD-RAM drives can be installed:

- DVD-RAM drive, FC 5751
- DVD-ROM drive, FC 2640

A logical partition running a supported release of the Linux® operating system requires the media enclosure and backplane feature code, and a DVD-ROM drive or DVD-RAM drive.

1.3.4 USB diskette drive

For today's administration tasks, an internal diskette drive is not state-of-the-art, but in some situations the external USB 1.44 MB diskette drive for p5-520 systems (FC 2591) is helpful. This super-slim-line and lightweight USB V2 attached diskette drive takes its power requirements from the USB port. A USB cable is provided. The drive can be attached to the integrated USB ports or to a USB adapter (FC 2738). A maximum of one USB diskette drive is supported per integrated controller/adapter. The same controller can share a USB mouse and keyboard.

1.3.5 I/O drawers

The p5-570 has six internal blind swap PCI-X slots: five are long slots and one is a short slot. The short PCI-X slot may also be used for the Remote I/O expansion card (FC 1800). If more PCI-X slots are needed, such as to extend the number of LPARs and partitions using Micro-Partitioning technology, up to 20 7311 Model D10, 7311 Model D11, or 7311 Model D20 I/O drawers can be attached.

7311 Model D10 I/O drawer

The 7311 Model D10 I/O drawer is supported if it is migrated from an existing system, but it cannot be ordered with the p5-570. The 7311 Model D10 is a 4U half-wide drawer that must be mounted in the rack enclosure (FC 7311) where two 7311 Model D10 I/O drawers could be mounted side-by-side. It features five hot-pluggable PCI-X slots and one standard hot-plug PCI slot with blind swap mechanism. (FC 4599 is the PCI Blind Swap Cassette Kit, Single Wide Adapter, Universal.) The 7311 Model D10 I/O drawer includes redundant concurrently maintainable power and cooling devices as default. The 7311 Model D10 I/O drawer does not slide out from the enclosure on rails, and therefore the IBM service representative must remove it for service.

The p5-570 system supports up to 20 7311 Model D10 drawers. A fully optioned system supports up to 124 PCI-X slots and up to 20 PCI slots (in full configuration, one PCI-X slot must be reserved for a Remote I/O expansion card).

The drawer has the following attributes:

- 4U rack-mount enclosure that can host one or two D10 drawers
- Six adapter slots
  - Five PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug
  - One PCI slot: 5 V, keyed, 33 MHz blind-swap hot-plug
- Default redundant hot-plug power and cooling devices
- Two RIO-2 and two SPCN ports

Note: The 7311 Model D10 I/O drawers require FC 6431 to ensure RIO-2 port capability, or an upgrade to RIO-2 is requested to support the connection to the p5-570 system.

7311 Model D10 I/O drawer physical package

Because the 7311 Model D10 I/O drawer must be mounted into the rack enclosure (FC 7311), these are the physical characteristics of one I/O drawer or two I/O drawers side-by-side:

- One 7311 Model D10 I/O drawer
  - Width: 223 mm (8.8 in)
  - Depth: 711 mm (28.0 in)
  - Height: 175 mm (6.9 in)
  - Weight: 19.6 kg (43 lb)
- Two I/O drawers in a 7311 rack-mounted enclosure have the following characteristics:
  - Width: 445 mm (17.5 in)
  - Depth: 711 mm (28.0 in)
  - Height: 175 mm (6.9 in)
  - Weight: 39.1 kg (86 lb)

Figure 1-2 on page 8 shows the different views of the 7311 Model D10 I/O drawer.

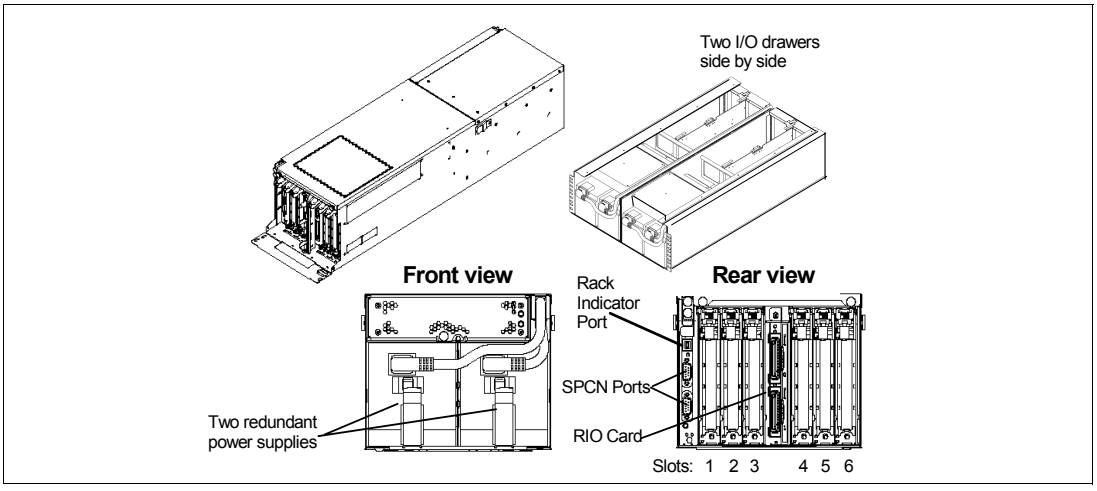


Figure 1-2 7311-D10 I/O drawer views

7311 Model D11 I/O drawer

The 7311 Model D11 I/O drawer is very similar to the 7311 Model D10, except that it features six long PCI-X slots and uses an improved blind-swap cassette design. Only the improved blind-swap cassettes are supported (FC 7862, for full-sized PCI cards), so the previous release of blind-swap cassettes, which were used in the 7311 Model D10 I/O drawer, are not supported.

Two 7311 Model D11 I/O drawers or two 7311 Model D10 I/O drawers fit side-by-side in the 4U enclosure (FC 7311) mounted in a 19-inch rack, such as the IBM 7014-T00 or 7014-T42.

The 7311 Model D11 I/O drawer offers a modular growth path for the p5-570 systems with increasing I/O requirements. A fully configured p5-570 supports 20 attached 7311 Model D11 I/O drawers. The combined system supports up to 144 PCI-X adapters. (In full configuration, Remote I/O expansion cards are required.)

The I/O drawer has the following attributes:

- 4U rack-mount enclosure (FC 7311) that can hold one or two D11 drawers
- Six PCI-X slots: 3.3 V, keyed, 133 MHz blind-swap hot-plug
- Default redundant hot-plug power and cooling devices
- Two RIO-2 and two SPCN ports

7311 Model D11 I/O drawer physical package

The I/O drawer enclosure has the same physical characteristics of the 7311 Model D10 I/O drawer; therefore the width, depth, height and weight dimensions are the same as described in “7311 Model D10 I/O drawer physical package” on page 8.

7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer is a 4U full-size drawer, which must be mounted in a rack. It features seven hot-pluggable PCI-X slots and optionally up to 12 hot-swappable disks arranged in two 6-packs. Redundant concurrently maintainable power and cooling is an optional feature (FC 6268). The 7311 Model D20 I/O drawer offers a modular growth path for the p5-570 systems with increasing I/O requirements. When a p5-570 is fully configured with 20 attached 7311 Model D20 drawers, the combined system supports up to 164 PCI-X adapters (in full configuration, a Remote I/O expansion card must be present, and 264 hot-swappable disks, for a total internal storage capacity of 38.7 TB using the 146.8 GB drive.

PCI-X and PCI cards are inserted into the slot from the top of the I/O drawer. The installed adapters are protected by plastic separators, which are designed to prevent grounding and damage when adding or removing adapters.

The drawer has the following attributes:

- 4U rack mount enclosure assembly
- Seven PCI-X slots: 3.3 V, keyed, 133 MHz hot-plug
- Two 6-pack hot-swappable SCSI devices
- Optional redundant hot-plug power
- Two RIO-2 and two SPCN ports

Note: The 7311 Model D20 I/O drawer initial order, or an existing 7311 Model D20 I/O drawer that is migrated from another pSeries system, must have the RIO-2 ports available (FC 6417).

7311 Model D20 I/O drawer physical package

The I/O drawer has the following physical characteristics:

- Width: 482 mm (19.0 in)
- Depth: 610 mm (24.0 in)
- Height: 178 mm (7.0 in)
- Weight: 45.9 kg (101 lb)

Figure 1-3 on page 10 shows the different views of the 7311-D20 I/O drawer.

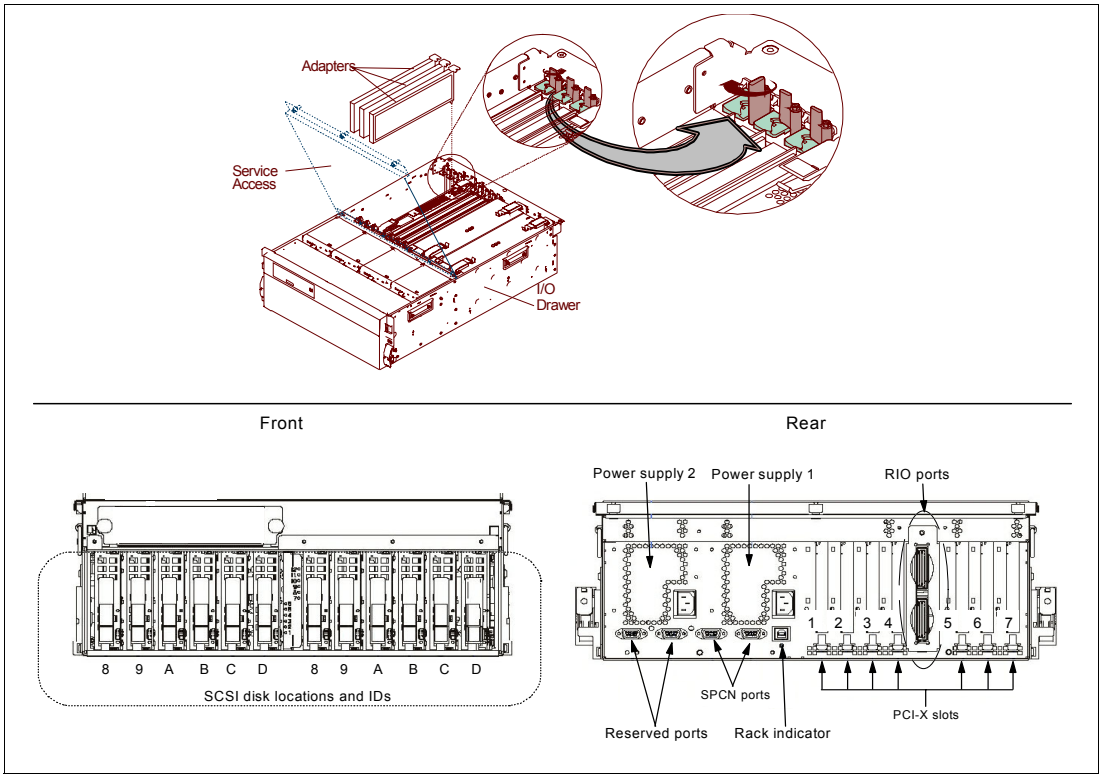


Figure 1-3 7311-D20 I/O drawer views

Note: The 7311 Model D10, and the 7311 Model D11, or the 7311 Model D20 I/O drawers are designed to be installed by an IBM service representative.

I/O drawers and usable PCI slots

The different I/O drawer model types can be intermixed on a single p5-570 server and within the same RIO-2 loop. Depending on the p5-570 system configuration, the maximum number of I/O drawers supported is different. Table 1-7 summarizes the maximum number of I/O drawers supported and the total number of PCI-X slots available when expansion consists of a single drawer type.

Table 1-7 Maximum number of I/O drawers supported and total number of PCI slots

p5-570 drawer/processors	Max number of I/O drawers	Total number of PCI-X slots		
		D10	D11	D20
1 drawer / 2-way	4	30[a]	30	34

p5-570 drawer/processors	Max number of I/O drawers	Total number of PCI-X slots		
		D10	D11	D20
1 drawer / 4-way	8	54[a]	54	62
2 drawers / 8-way	12	84[a]	84	96
3 drawers / 12-way	16	114[a]	114	130
4 drawers / 16-way	20	144[ab]	144[b]	164

a. One slot per drawer is PCI.
b. One slot is reserved for the Remote I/O expansion card.

1.3.6 Hardware Management Console models

The Hardware Management Console (HMC) provides a set of functions that is necessary to manage the p5-570 system when LPAR, Capacity on Demand without reboot, inventory and microcode management, and remote power control functions are needed. These functions include the handling of the partition profiles that define the processor, memory, and I/O resources that are allocated to an individual partition.

The 7310 Model CR2 and the 7310 Model C03 HMC are specifically for POWER5 processor-based systems. However, an existing 7315 Model CR2 and the 7315 Model C03 (POWER4 processor-based systems HMC) can be converted for POWER5 processor-based system use when it is loaded with the HMC software that is required for POWER5 processor-based systems (FC 0961).

POWER5 processor-based system HMCs require Ethernet connectivity. Ensure that sufficient Ethernet adapters are available to enable public and private networks if you need both. The 7310 Model C03 is a desktop model with only one native 10/100/1000 Ethernet port, but three additional PCI slots. The 7310 Model CR2 is a 1U, 19-inch rack-mountable drawer that has two native Ethernet ports and two additional PCI slots.

When an HMC is connected to the p5-570, the p5-570 integrated serial ports are disabled. If you need serial connections, for example non-Ethernet HACMP™ heartbeat, you must provide an async adapter.

Note: It is not possible to connect POWER4 and POWER5 processor-based systems to the same HMC simultaneously.

1.4 Value Paks

Value Paks are a new offering that is available on an initial order only. They provide a predefined configuration that is designed to meet typical customer requirements. Activations are available when a system order satisfies specific configuration requirements for memory, disk drives, and processors. When a Value Pak is ordered, you can select additional features. Customers can configure systems with 2 to 16 processors and 2 to 16 processor activations. For each paid processor activation, the customer is entitled to one processor activation at no additional charge if the following requirements are met:

- The system must have at least two disk drives of at least 73.4 GB each.
- There must be at least 2 GB of memory installed for each activated processor, as described in Table 1-8.

Table 1-8 Value Pak configuration

Value Paks	Processors, FC	Building blocks	Memory (MB), FC	Disk
1.5 GHz	2-way, 7834 x 1	1	4096 MB, 4452 x 2	2x73.6 GB (FC 3274)
1.5 GHz	4-way, 7834 x 2	1	8192 MB, 4452 x 4	2x73.6 GB (FC 3274)
1.5 GHz	8-way, 7834 x 4	2	16384 MB, 4452 x 8	2x73.6 GB (FC 3274)
1.65 GHz	2-way, 7830 x 1	1	4096 MB, 4452 x 2	2x73.6 GB (FC 3274)
1.65 GHz	4-way, 7830 x 2	1	8192 MB, 4452 x 4	2x73.6 GB (FC 3274)
1.65 GHz	8-way, 7830 x 4	2	16384 MB, 4452 x 8	2x73.6 GB (FC 3274)
1.9 GHz	4-way, 7832 x 2	1	8192 MB, 4452 x 4	2x73.6 GB (FC 3274)
1.9 GHz	8-way, 7832 x 4	2	16384 MB, 4452 x 8	2x73.6 GB (FC 3274)
1.9 GHz	16-way, 7832 x 8	4	32768 MB, 4452 x 16	2x73.6 GB (FC 3274)
1.9 GHz	4-way, 7833 x 2	1	8192 MB, 7892 x 4	2x73.6 GB (FC 3274)
1.9 GHz	8-way, 7833 x 4	2	16384 MB, 7892 x 8	2x73.6 GB (FC 3274)
1.9 GHz	16-way, 7833 x 8	4	32768 MB, 7892 x 16	2x73.6 GB (FC 3274)

1.5 Model type conversion

Customers who own an IBM pSeries 650 system may convert their system to an IBM ℯserver p5 570 system. System hardware for the new model will consist of one or more drawers to replace the old system chassis. Supported features from the old model will be transferred to the new system.

Model conversions allow any valid number of processors in the new model, regardless of the number of processors in the old model. The valid processor quantities are 2, 4, 8, 12, and 16. Any supported memory and disk drive features that are transferred from the previous system can be counted toward the requirements for no additional charge processor activations, using the same Value Pak rules that apply to new system orders. The p5-570 primary building block retains the serial number of the older IBM pSeries 650. Upon completion, the previous IBM pSeries 650 system will be returned to IBM. Parts that are removed or replaced become the property of IBM. Supported features cannot be ordered on the converted model, but they can be left on or removed from the converted model.

1.6 System racks

The Enterprise Rack Models T00 and T42, which are 19 inches wide, are general-use racks with IBM ℯserver p5, pSeries, and RS/6000 rack-based or rack drawer–based systems. The racks provide increased capacity, greater flexibility, and improved floor space utilization.

The p5-570 uses a 4U rack-mounted server drawer as its basic building block. If a rack is ordered without a door, an enhanced front trim kit (FC 6246 for the T00, or FC 6247 for the T42) provides the additional clearance that is required for the p5-570 bezel.

If an IBM ℯserver p5 system is to be installed in a non-IBM rack or cabinet, you should ensure that the rack conforms to the EIA[3] standard EIA-310-D. (See 1.6.6, “OEM rack” on page 16.)

Note: It is the customer's responsibility to ensure that the installation of the drawer in the preferred rack or cabinet results in a configuration that is stable, serviceable, safe, and compatible with the drawer requirements for power, cooling, cable management, weight, and rail security.

1.6.1 IBM RS/6000 7014 Model T00 Enterprise Rack

The 1.8-meter (71-inch) Model T00 is compatible with past and current p5, pSeries, and RS/6000 racks, and is designed for use in all situations that previously used the older rack models R00 and S00. The T00 rack has the following features:

- 36 EIA units (36U) of usable space.
- Optional removable side panels.
- Optional highly perforated front door.
- Optional side-to-side mounting hardware for joining multiple racks.
- Standard black or optional white color in OEM format.
- Increased power distribution and weight capacity.
- Optional reinforced (ruggedized) rack feature (FC 6080) provides added earthquake protection with modular rear brace, concrete floor bolt-down hardware, and bolt-in steel front filler panels.
- Support for both AC and DC configurations.
- DC rack height is increased to 1926 mm (75.8 in) if a power distribution panel is fixed to the top of the rack.
- Up to four Power Distribution Units (PDUs) can be mounted in the proper bays, but others can fit inside the rack. (See 1.6.3, "AC Power Distribution Unit and rack content" on page 14.)
- An optional rack status beacon (FC 4690). This beacon is designed to be placed on top of a rack and cabled to servers, such as a p5-570, and to other components, such as a 7311 I/O drawer, inside the rack. Servers can be programmed to illuminate the beacon in response to a detected problem or changes in system status.
- A rack status beacon junction box (FC 4693) should be used to connect multiple servers and I/O drawers to the beacon. This feature provides six input connectors and one output connector for the rack. To connect the servers or other components to the junction box or the junction box to the rack, status beacon cables (FC 4691) are necessary. Multiple junction boxes can be linked in a series using daisy chain cables (FC 4692).
- Weight:

  T00 base empty rack: 244 kg (535 lb)

  T00 full rack: 816 kg (1795 lb)

[3] Electronic Industries Alliance (EIA). Accredited by American National Standards Institute (ANSI), EIA provides a forum for industry to develop standards and publications throughout the electronics and high-tech industries.

1.6.2 IBM RS/6000 7014 Model T42 Enterprise Rack

The 2.0-meter (79.3-in) Model T42 is the rack that will address the special requirements of customers who want a tall enclosure to house the maximum amount of equipment in the smallest possible floor space. The features that differ in the Model T42 rack from the Model T00 include the following:

- 42 EIA units (42U) of usable space
- AC power support only
- Weight:
  - T42 base empty rack: 261 kg (575 lb)
  - T42 full rack: 930 kg (2045 lb)

1.6.3 AC Power Distribution Unit and rack content

For rack models T00 and T42 nine-outlet PDUs are available.

PDUs with nine outlets (FC 9176, 9177, 9178, 7176, 7177, and 7178) are available. A T42 rack that is configured for the maximum number of power outlets would have six PDUs (two mounted horizontally requiring 2U of rack space), for a total of 54 power outlets.

The p5-570 can be connected to any PDU that is available for the 7014-T00 or 7014-T42 racks.

For detailed power cord requirements and power cord feature codes, see the publication *Site and Hardware Planning Information*, SA38-0508. An online copy can be found at:

`http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/`

The first four PDUs that are ordered for a rack are mounted vertically in the sides of the rack, occupying all the four available PDU bays. Any additional PDU will be mounted horizontally in the rear of the rack and will occupy 1U of rack space.

Note: Each p5-570 building block, or a system drawer to be mounted in the rack, requires two power cords, which are not included in the base system order.

Universal PDU (FC 7188) and the optional Universal PDU to be mounted horizontally in the rack (FC 9188) will be available on December 31, 2004, supporting a wide range of country requirements and electrical power specifications. Each Universal PDU provides 12 C13 power outlets for use within a 7014-T00 or 7014-T42 rack, compared to nine C13 power outlets provided by the FC 7176 or FC 7177 PDUs. Nine different power cord features can be used to connect the PDU to a wall power outlet. Each power cord provides the unique design characteristics for the different power requirements. To match new power requirements and save previous investments, these power cords could be requested with an initial order of the rack, or with a later upgrade of the rack features.

1.6.4 Rack-mounting rules for p5-570 and I/O drawers

The primary rules that should be followed when mounting the p5-570 into a rack are:

- The p5-570 is designed to be placed at any location in the rack. For rack stability, it is advisable to start filling a rack from the bottom.

  For p5-570 configurations with 2, 3, or 4 drawers, all drawers must be installed together in the same rack, in a contiguous space of 8U, 12U, or 16U within the rack.

- Any remaining space in the rack can be used to install other systems or peripherals, provided that the maximum permissible weight of the rack is not exceeded and the installation rules for these devices are followed.
- Before placing a p5-570 into the service position, it is essential that the rack manufacturer's safety instructions regarding rack stability have been followed.
- Special consideration must be taken to avoid a flange on the top of the rack to clear the front bezel.

Depending on current implementation and future enhancements of additional 7311 Model D10, 7311 Model D11, and 7311 Model D20 I/O drawers connected to the p5-570 or single installed p5-570 systems, Table 1-9 on page 15 shows examples of the minimum and maximum configurations for p5-570 systems and attached 7311 Model D20 I/O drawers. (7311 Model D10 is a half drawer, so a 7311 Model D20 takes the place of two 7311 Model D10 mounted side-by-side.)

Table 1-9 Minimum and maximum configurations for p5-570s and 7311-D20s

Rack	Stand-alone p5-570s	One p5-570, one 7311-D20	One p5-570, four 7311-D20s	One p5-570, eight 7311-D20s
7014-T00	9	4	1	1
7014-T42	10	5	2	1

The front trim kit consists of three steel parts that snap on to the rack at the left, right, and top edges. They are painted to match the overall rack color and designed to present an attractive, finished appearance. Each rack must be ordered with either a front door or a front trim kit, but not both. The front trim kit is offered with two feature codes, one for the 7014-T00 and the other for the 7014-T42 racks (FC 6246, FC 6247). The FC 6246 and FC 6247 replaces the earlier trim kits. Either feature may be installed on a 7014-T42 rack, unless the rack contains a p5-570 system. The p5-570 is not compatible with the FC 6081 trim kit.

1.6.5 Additional options for rack

The intention of this section is to highlight some solutions available to provide a single point of management for a complex environment, such as with several p5-570 systems inside a single T00 or T42 rack. In addition, this section highlights the IBM 7212 Model 102 TotalStorage® device enclosure, to connect an external tape drive solution to the p5-570.

Flat panel display options

The IBM 7316-TF3 Flat Panel Console Kit may be installed in the system rack. This 1U console uses a 15-inch thin film transistor (TFT) LCD with a viewable area of 304.1 mm x 228.1 mm and a 1024 x 768 resolution. The 7316-TF3 Flat Panel Console Kit has the following attributes:

- Flat panel color monitor.
- Rack tray for keyboard, monitor, and optional VGA switch with mounting brackets.
- IBM Space Saver 2, a 14.5-inch keyboard that mounts in the rack keyboard tray and is available as a feature in 16 language configurations. (The track point mouse is integrated into the keyboard.)

Note: It is recommended that you have the 7316-TF3 installed between EIA 20 and- 25 of the rack for ease of use. The 7316-TF3 or any other graphics monitor requires that a POWER GXT135P graphics accelerator (FC 2848 or FC 2849) is installed in the server.

Hardware Management Console 7310 Model CR2

The 7310 Model CR2 is a 1U, 19-inch rack-mountable drawer supported in the 7014 Model T00 and T42 racks. The 7310 Model CR2 provides one serial port, two integrated Ethernet ports, and two additional PCI slots. The HMC 7310 Model CR2 has USB ports to connect USB keyboard and mouse devices.

Note: The HMC serial port can be used for external modem attachment if the Service Agent call-home function is implemented, and the Ethernet ports are used to communicate to the service processor in p5-570 systems. An Ethernet cable (FC 7801 or 7802) is required to attach the HMC to the p5-570 system it controls.

IBM 7212 Model 102 TotalStorage Storage device enclosure

The IBM 7212 Model 102 is designed to provide efficient and convenient storage expansion capabilities for selected IBM @server p5, pSeries, and RS/6000 servers. The IBM 7212 Model 102 is a 1U rack-mountable option that can be installed in a standard 19-inch rack using an optional rack-mount hardware feature kit. The 7212 Model 102 has two bays that can accommodate any of the following storage drive features:

- DDS Gen 5 DAT72 Tape Drive provides physical storage capacity of 36 GB (72 GB with 2:1 compression) per data cartridge.
- VXA-2 Tape Drive provides a media capacity of up to 80 GB (160 GB with 2:1 compression) physical data storage capacity per cartridge.
- Digital Data Storage (DDS-4) tape drive with 20 GB native data capacity per tape cartridge and a native physical data transfer rate of up to 3 MBps, uses 2:1 compression so that a single tape cartridge can store up to 40 GB of data.
- DVD-ROM drive is a 5.25-inch, half-high device. It can read 640 MB CD-ROM and 4.7 GB DVD-RAM media. It can be used for Alternate IPL[4] (IBM-distributed CD-ROM media only) and program distribution.
- DVD-RAM drive with up to 2.7 MBps throughput. Using 3:1 compression, a single disk can store up to 28 GB of data. Supported DVD disk native capacities on a single DVD-RAM disk are as follows: up to 2.6 GB, 4.7 GB, 5.2 GB, and 9.4 GB.

1.6.6 OEM rack

The p5-570 can be installed in a suitable OEM rack, as long as the rack conforms to the EIA-310-D standard and careful consideration is given to the additional 2.75 inches required on the left front side for interconnect cables when multiple drawers installed and connected. This standard is published by the Electrical Industries Alliance, and a summary of this standard is available in the publication *Site and Hardware Planning Information*, SA38-0508.

The key points that are mentioned in this standard are:

- Any rack that is used must be capable of supporting 15.9 kg (35 lb) per EIA unit (44.5 mm [1.75 in]) of rack height.
- To ensure proper rail alignment, the rack must have mounting flanges that are at least 494 mm (19.45 in) across the width of the rack and 719 mm (28.3 in) between the front and rear rack flanges.
- It may be necessary to supply additional hardware, such as fasteners, for use in some manufacturer's racks.

[4] Initial program load

1.7 Statement of direction

IBM plans to extend the capabilities of the IBM ℯserver p5 product line by introducing support for the i5/OS™ operating system. This support is planned for selected IBM ℯserver p5 570 and future high-end IBM ℯserver p5 models. i5/OS support will provide additional flexibility for large-scale server consolidation where AIX 5L™ and/or Linux is the primary operating system. i5/OS support will be limited to one processor on selected p5-570 models. This capability is planned to be available in the first half of 2005.

All statements regarding IBM's future direction and intent are subject to change or withdrawal without notice, and represent goals and objectives only.

2

Architecture and technical overview

This chapter discusses the overall system architecture represented by Figure 2-1, with its major components described in the following sections. The bandwidths that are provided throughout the section are theoretical maximums used for reference. You should always obtain real-world performance measurements using production workloads.

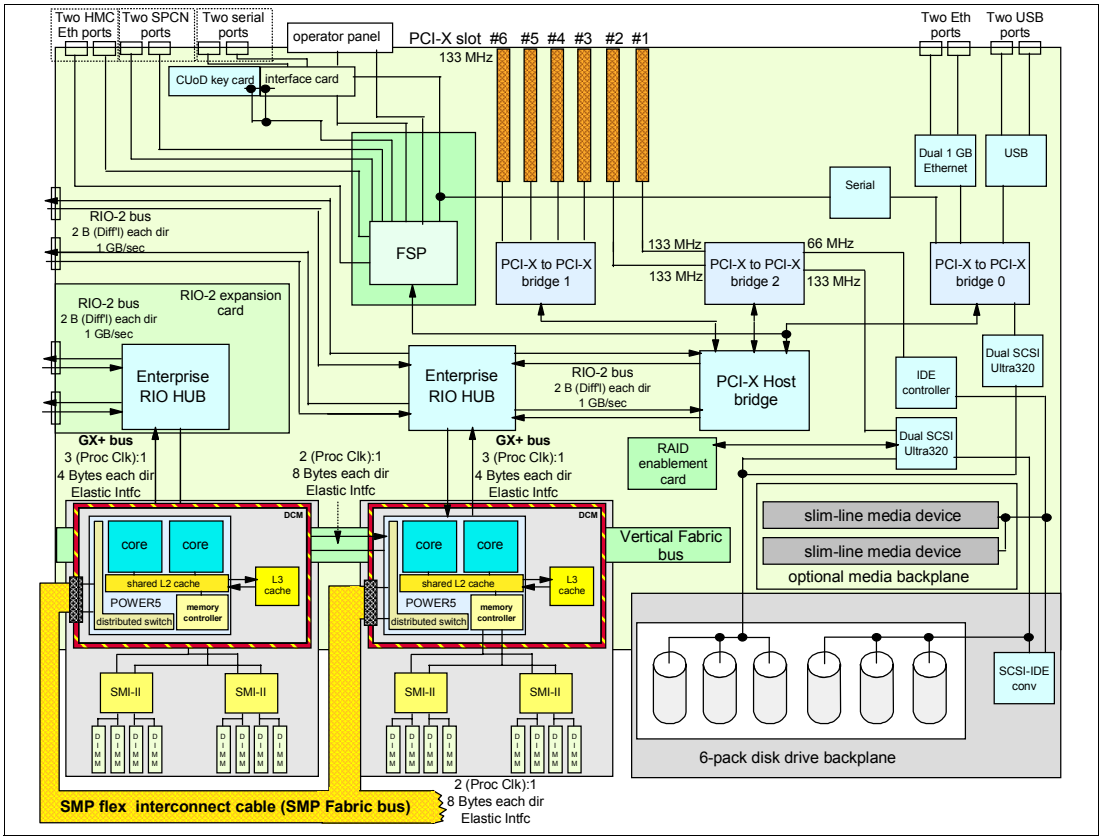


Figure 2-1 p5-570 logic data flow

2.1 The POWER5 chip

The POWER5 chip features single-threaded and multi-threaded execution, providing higher performance in the single-threaded mode than its POWER4 predecessor provides at equivalent frequencies. POWER5 maintains both binary and architectural compatibility with existing POWER4 systems to ensure that binaries continue executing properly and that all application optimizations carry forward to newer systems. POWER5 provides additional enhancements such as virtualization, improved reliability, availability, and serviceability at both chip and system levels, and it has been designed to support speeds up to 3 GHz.

Figure 2-2 shows the high-level structures of POWER4 and POWER5 processor-based systems. The POWER4 scales up to a 32-way symmetric multiprocessor. Going beyond 32 processors increases interprocessor communication, resulting in higher traffic on the interconnection fabric bus. This can cause greater contention and negatively affect system scalability.

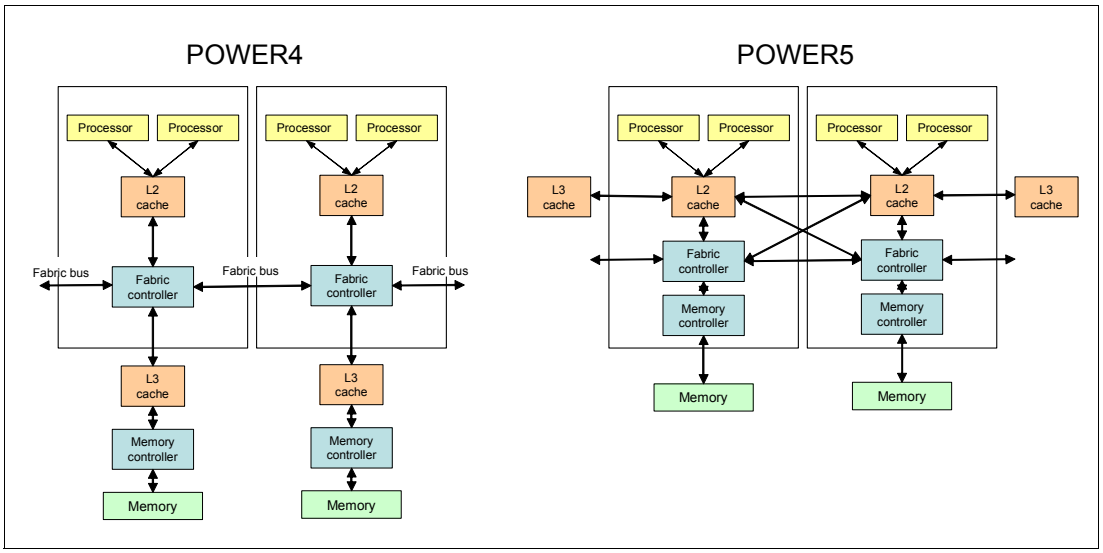


Figure 2-2 POWER4 and POWER5 system structures

Moving the L3 cache provides significantly more cache on the processor side than was available previously, thus reducing traffic on the fabric bus and enabling POWER5 processor-based systems to scale to higher levels of symmetric multiprocessing. The POWER5 supports a 1.9 MB on-chip L2 cache, implemented as three identical slices with separate controllers for each. Either processor core can independently access each L2 controller. The L3 cache, with a capacity of 36 MB, operates as a backdoor with separate buses for reads and writes that operate at half processor speed.

Because of the higher transistor density of the POWER5 0.13-µm technology, it was possible to move the memory controller on-chip and eliminate a chip that was previously needed for the memory controller function. These changes in the POWER5 processor also have the significant side benefits of reducing latency to the L3 cache and main memory, as well as reducing the number of chips that are necessary to build a system.

The POWER5 processor supports the 64-bit PowerPC® architecture. A single die contains two identical processor cores, each supporting two logical threads. This architecture makes the chip appear as a four-way symmetric multiprocessor to the operating system. The POWER5 processor core has been designed to support both enhanced simultaneous multi-threading (SMT) and single-threaded (ST) operation modes.

2.1.1 Simultaneous multi-threading

As a permanent requirement for performance improvements at the application level, simultaneous multi-threading (SMT) functionality is embedded in the POWER5 chip technology. Developers are familiar with process-level parallelism (multi-tasking) and thread-level parallelism (multi-threads). SMT is the next stage of processor saturation for throughput-oriented applications to introduce the method of instruction group-level parallelism to support multiple pipelines to the processor. The instruction groups are chosen from different hardware threads belonging to a single OS image.

SMT is activated by default when an OS that supports it is loaded. On a 2-way POWER5 processor based system, the operating system discovers the available processors as a 4-way system. To achieve a higher performance level, SMT is also applicable in Micro-Partitioning, capped or uncapped, and dedicated partition environments (2.9, “Virtualization” on page 38).

Simultaneous multi-threading is supported on POWER5 processor-based systems running AIX 5L V5.3 or Linux-based systems at a required 2.6 kernel. AIX provides the **smtctl** command that turns SMT on and off without subsequent reboot. For Linux, an additional boot option must be set to activate SMT after a reboot.

The SMT mode maximizes the usage of the execution units. In the POWER5 chip, more rename registers have been introduced (for Floating Point operation, rename registers are increased to 120), that are essential for out-of-order execution and vital for the SMT.

Enhanced SMT features

To improve SMT performance for various workload mixes and provide robust quality of service, POWER5 provides two features:

- Dynamic resource balancing
  - The objective of dynamic resource balancing is to ensure that the two threads executing on the same processor flow smoothly through the system.
  - Depending on the situation, the POWER5 processor resource balancing logic has a different thread throttling mechanism.
- Adjustable thread priority
  - Adjustable thread priority lets software determine when one thread should have a greater (or lesser) share of execution resources.
  - The POWER5 supports eight software-controlled priority levels for each thread.

ST operation

Not all applications benefit from SMT. Having threads executing on the same processor does not increase the performance of applications with execution unit limited performance or applications that consume all of the chip’s memory bandwidth. For this reason, the POWER5 processor supports the ST execution mode. In this mode, the POWER5 processor gives all of the physical resources to the active thread, enabling it to achieve higher performance than a POWER4 processor-based system at equivalent frequencies. Highly optimized scientific codes are one example where ST operation is ideal.

2.1.2 Dynamic power management

In current CMOS[1] technologies, chip power is one of the most important design parameters. With the introduction of SMT, more instructions execute per cycle per processor core, thus increasing the core’s and the chip’s total switching power. To reduce switching power,

[1] complementary metal oxide semiconductor

POWER5 chips extensively use a fine-grained, dynamic clock-gating mechanism. This mechanism gates off clocks to a local clock buffer if dynamic power management logic knows that the set of latches that are driven by the buffer will not be used in the next cycle. This allows substantial power saving with no performance impact. In every cycle, the dynamic power management logic determines whether a local clock buffer that drives a set of latches can be clock-gated in the next cycle.

In addition to the switching power, leakage power has become a performance limiter. To reduce leakage power, the POWER5 chip uses transistors with low threshold voltage only in critical paths. The POWER5 chip also has a low-power mode, enabled when the system software instructs the hardware to execute both threads at the lowest available priority. In low power mode, instructions dispatch once every 32 cycles at most, further reducing switching power. The POWER5 chip uses this mode only when there is no ready task to run on either thread.

2.1.3 The POWER chip evolution

The p5-570 system complies with the RS/6000 platform architecture, which is an evolution of the PowerPC Common Hardware Reference Platform (CHRP) specifications. Figure 2-3 shows the POWER chip evolution.

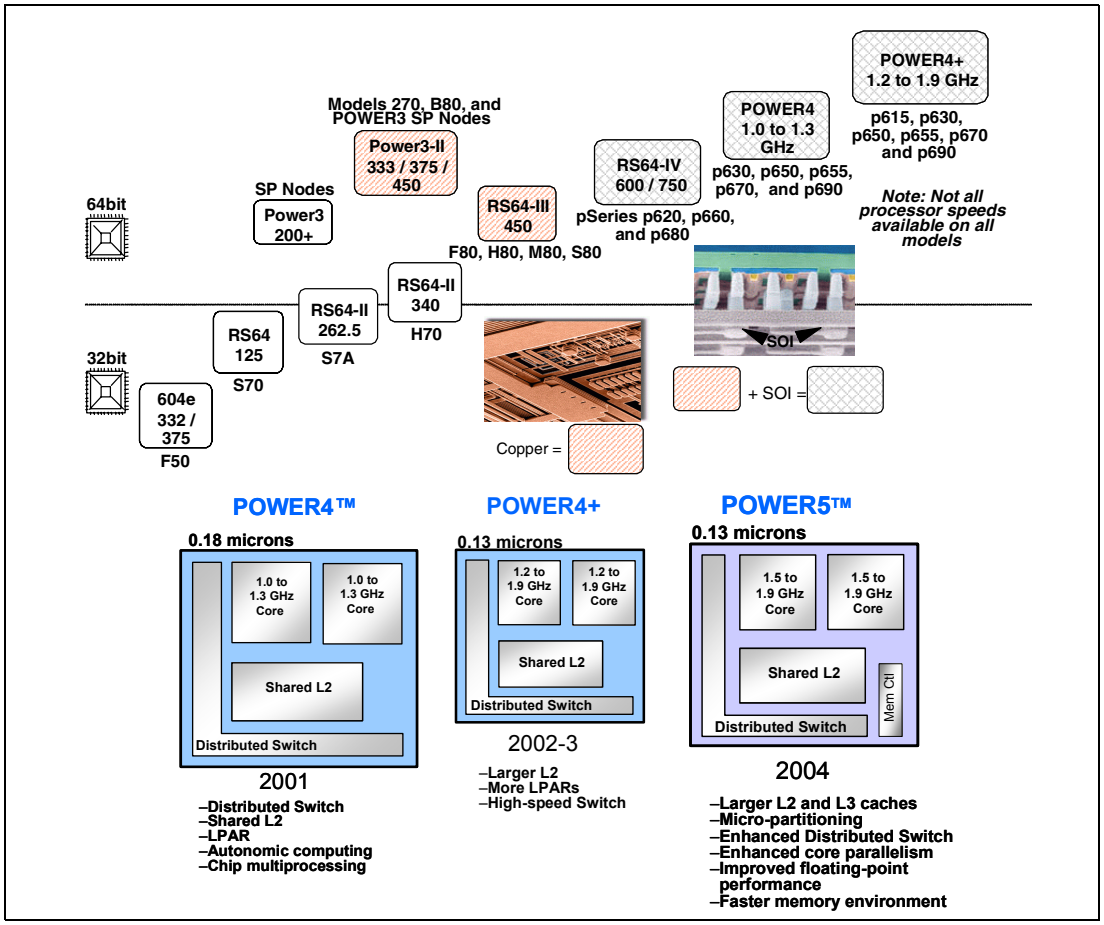


Figure 2-3 The POWER chip evolution

2.1.4 CMOS, copper, and SOI technology

The POWER5 processor design is a result of a close collaboration between *IBM Systems and Technology Group* and *IBM Microelectronics technologies* that enables IBM *@*server p5 systems to give customers improved performance, reduced power consumption, and decreased IT footprint size through logical partitioning. The POWER5 processor chip takes advantage of IBM leadership technology. It is made using IBM 0.13-µm-lithography CMOS. The POWER5 processor also uses copper and Silicon-on-Insulator (SOI) technology to allow a higher operating frequency for improved performance, yet with reduced power consumption and improved reliability compared to processors not using this technology.

2.2 Processor cards

In the p5-570 system, the POWER5 chip has been packaged with the L3 cache chip into a cost-effective Dual Chip Module (DCM) package. The storage structure for the POWER5 processor chip is a distributed memory architecture that provides high-memory bandwidth. Each processor can address all memory and sees a single shared memory resource. As such, a single DCM and its associated L3 cache and memory are packaged on a single processor card. Access to memory behind another processor is accomplished through the fabric buses. The p5-570 supports up to two processor cards (each card is a 2-way) in any building block. Each processor card has a single DCM containing a POWER5 processor chip and a 36 MB L3 module. I/O connects to the Central Electronic Complex (CEC) subsystem using the GX+ bus. Each DCM provides a single GX+ bus for a total system capability of two GX+ buses. The GX+ bus provides an interface to a single device such as the RIO-2 buses.

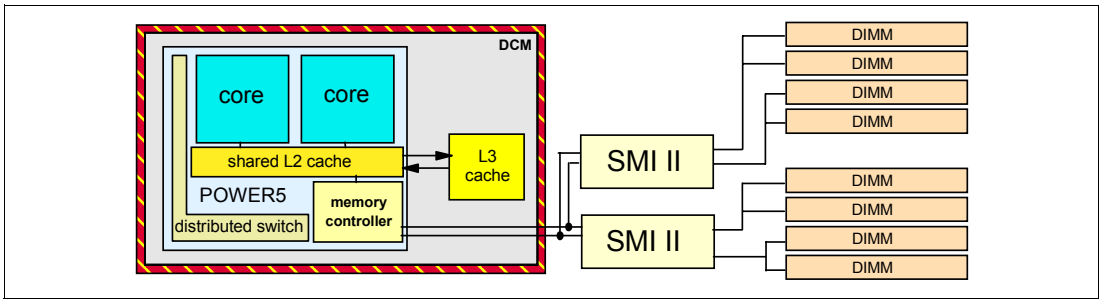


Figure 2-4 p5-570 DCM diagram

Each processor card contains a single DCM, as well as the local memory storage subsystem for that DCM. The processor card also contains LEDs for each FRU[2] on the CPU card including the CPU card itself. Figure 2-5 shows a processor card layout view.

[2] field replacement unit

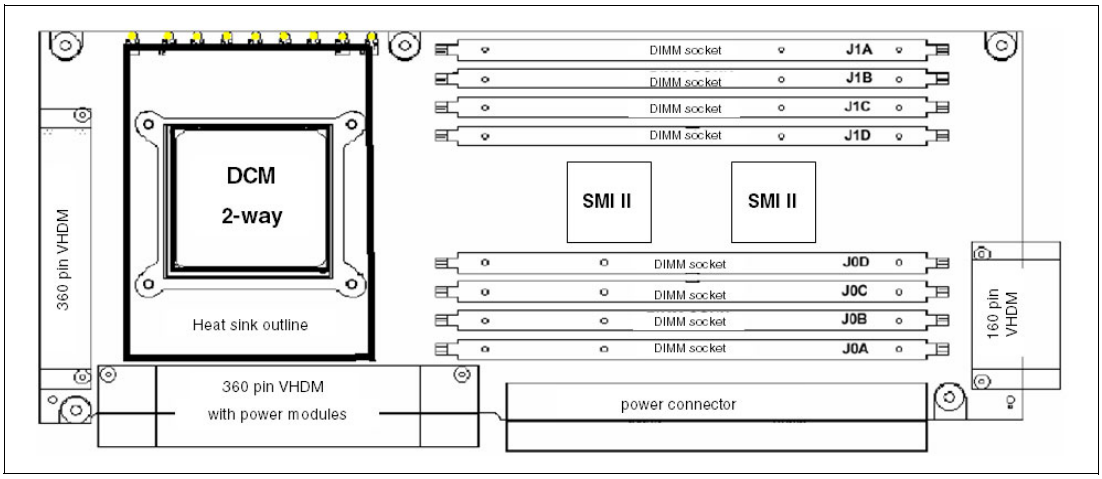


Figure 2-5 Processor card with DDR1 memory socket layout view

There are two system backplanes in the p5-570 system. A GX+ bus planar, which docks vertically into the system planar, is always present in the system. The processor cards dock directly into this backplane from the front. A horizontal backplane exists below the CPU cards that is co-planar with the I/O backplane. This backplane routes the vertical fabric bus between the processor cards. This backplane is also used for power distribution from the CPU regulators that are housed next to the processor cards. (See Figure 1-1 on page 4.)

2.2.1 Processor drawer interconnect cables

In combined systems that are made of more than one p5-570 building block, the connection between processor cards in different building blocks is provided with a processor drawer interconnect cable. Different processor drawer interconnect cables are required for the different numbers of p5-570 building blocks that a combined system can be made of, as shown in Figure 2-6.

Because of the redundancy and fault recovery built-in to the system interconnects, a drawer failure does not represent a system failure. Once a problem is isolated and repaired, a system reboot may be required to reestablish full bus speed, if the failure was specific to the interconnects.

The SMP fabric bus that connects the processors of separate p5-570 building blocks is routed on the interconnect cable that is routed external to the building blocks. The flexible cable attaches directly to the processor cards, at the front of the p5-570 building block, and is routed behind the front covers (bezels) of the p5-570 building blocks. There is an optimized cable for each drawer configuration. The Figure 2-6 illustrates the logical fabric bus connections between the drawers, and shows the additional space required left of the bezels for rack installation.

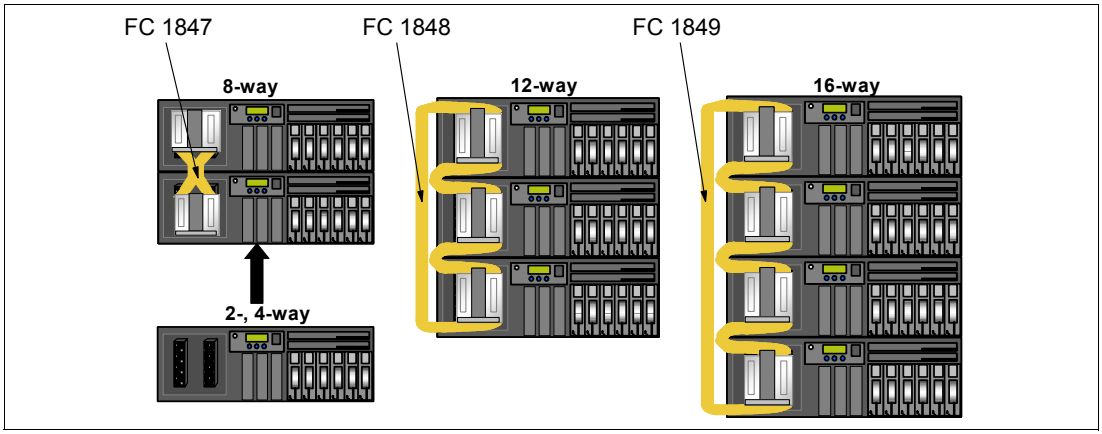


Figure 2-6 Logical p5-570 building blocks connection

2.2.2 Processor clock rate

The p5-570 system features base 2-way, 4-way, 8-way, 12-way, and 16-way configurations with the POWER5 processor running at 1.5 GHz, 1.65 GHz, and 1.9 GHz. The processor card running at 1.9 GHz is available with two feature codes to support the DDR1 and DDR2 memory technology, which require two different DIMM sockets.

Note: Any system made of more than one processor card must have all processor cards running at the same speed.

To determine the processor characteristics on a running system, use one of the following commands:

`lsattr -El proc`*X* Where *X* is the number of the processor; for example, proc0 is the first processor in the system. The output from the command[3] would be similar to this:

```
type powerPC_POWER5     Processor type        False
frequency 165600000     Processor Speed       False
smt_enabled true        Processor SMT enabled False
smt_threads 2           Processor SMT threads False
state enable            Processor state       False
```

(False, as used in this output, signifies that the value cannot be changed through an AIX command interface.)

`pmcycles -m` This command (AIX 5L Version 5.3 and later) uses the performance monitor cycle counter and the processor real-time clock to measure the actual processor clock speed in MHz. This is the output of a 2-way p5-550 running at 1.65 GHz system:

```
Cpu 0 runs at 1656 MHz
Cpu 1 runs at 1656 MHz
```

Note: The **`pmcycles`** command is part of the bos.pmapi fileset. First check whether that component is installed by using the **`lslpp -l bos.pmapi`** command.

[3] The output of the **`lsattr`** command has been expanded with AIX 5L to include the processor clock rate.

2.3 Memory subsystem

The p5-570 memory controller is internal to the POWER5 chip. It interfaces to either two (DDR1) or four (DDR2) SMI-II buffer chips and 8 pluggable DIMMs per processor card, as described in 2.2, “Processor cards” on page 23. The minimum memory for a p5-570 processor-based system is 2 GB. The maximum installable memory is 512 GB (using DDR1 memory DIMM technology). The p5-570 total memory depends on the number of available processor cards. Figure 2-7 shows memory slot availability.

2.3.1 Memory placement rules

The memory features that are available for the p5-570 at the time of writing are listed in 1.3.2, “Memory features” on page 6. Each memory feature consists of four DIMMs, or quad, and must be installed according to Figure 2-7. The first quad slots are J0A, J1A, J0C, and J1C. For the second quad, the slots are J0B, J1B, J0D, and J1D.

Note: A quad must consist of a single feature (that is, made of four identical DIMMs). Mixing DIMM capacities in a quad is not supported.

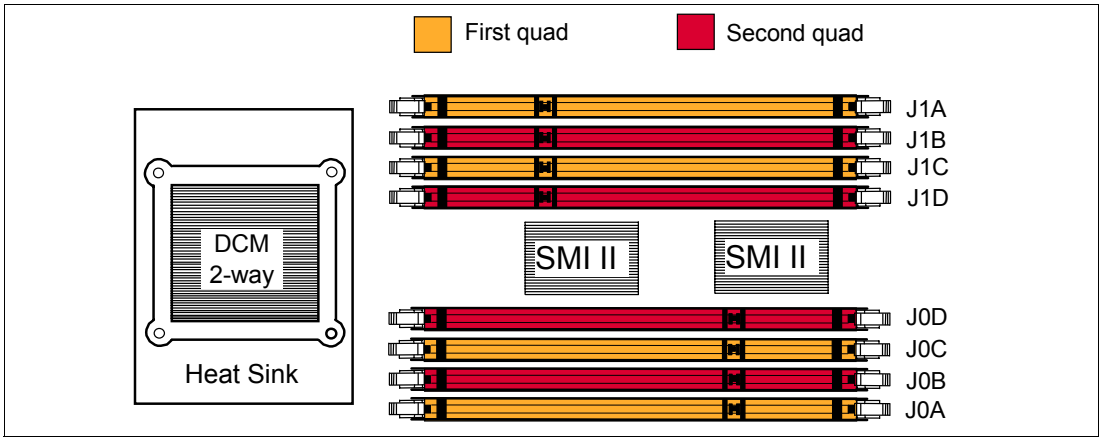


Figure 2-7 Memory placement for the p5-570, DDR1 card

2.3.2 Memory restriction

The p5-570 does not support OEM memory, and there is no exception to this rule. OEM memory is not certified for use in pSeries and in the IBM @server p5 systems. If the p5-570 is populated with OEM memory, you could experience unexpected and unpredictable behavior.

All IBM memory is identified by an IBM logo and a white label printed with a barcode on top and an alphanumeric string on the bottom, created according to the rule reported in Figure 2-8.

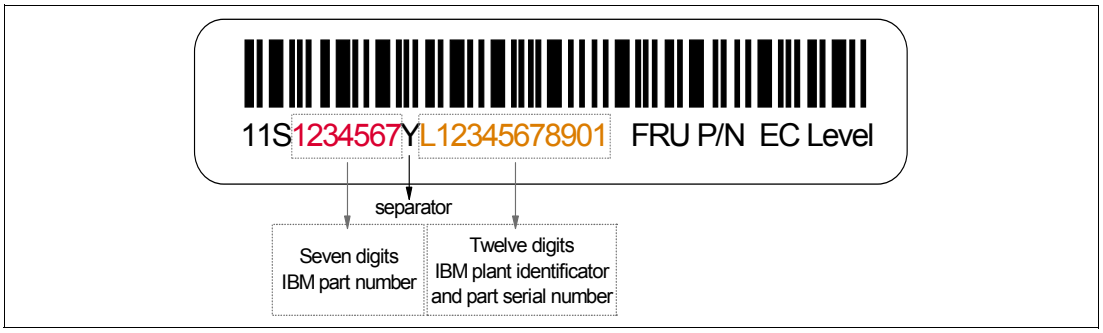


Figure 2-8 IBM memory certification label

Sometimes OEM vendors attach a label to their DIMMs that reports the IBM memory part number but not the barcode or the alphanumeric string.

In case of system failure caused by OEM memory installed in the system, the first thing to do is to replace the suspected memory with IBM memory, then check whether the problem is corrected. Contact your IBM representative for further assistance if needed.

2.3.3 Memory throughput

The memory subsystem throughput is based on the speed of the memory, not the speed of the processor. An elastic interface, contained in the POWER5 chip, buffers reads and writes to and from memory and the processor. On DDR1 cards, there are two SMIs, each with a single 8 byte read and 2 byte write DDR bus to the processor on each processor card. A DDR bus allows double reads or writes per clock cycle. If 266 MHz memory is installed, the throughput is (16 x 2 x 266.5) + (4 x 2 x 266.5) or 10660 MB/second or 10.41 GB/second per processor card. For a building block with two processor cards, this value is doubled, or 20.82 GB/second.

DDR2 processor cards contain an additional set of two SMIs to manage the increased throughput. However in this configuration the paths are 4 bytes for read operations and 2 bytes for write. Therefore the throughput is (4 + 2) * 4 * 1066 = 24.98 GB/s or 49.96 GB/s for a 4-way node. These values are maximum theoretical throughputs for comparison purposes only.

The POWER5 processor's integrated memory controller further reduces latency over the previous outboard controller on POWER4 systems to the SMI chips by requiring fewer cycles in order to set up memory addressing in the hardware.

2.4 System buses

The following sections provide additional information related to the internal buses.

2.4.1 RIO-2 buses and GX+ card

Each DCM provides a GX+ bus that is used to connect to an I/O subsystem or Fabric Interface card. In a p5-570 drawer, there are two GX+ buses, one from each processor card. Each p5-570 has one GX slot with a single GX+ bus. The GX+ slot is not active unless the second processor card is installed. It is not required for CUoD processor cards to be activated in order for the associated GX+ bus to be active. The p5-570 provides two external RIO-2 ports, which can operate up to 1 GHz. An add-in GX+ adapter card (Remote I/O expansion card, FC 1800) adds two more RIO-2 ports. When this card is installed, PCI adapter slot 6

must remain empty. All GX+ cards are hot-pluggable. The RIO-2 ports are used for I/O expansion to external I/O drawers. The supported I/O drawers are the 7311 Model D10, 7311 Model D11, and 7311 Model D20.

The Remote I/O expansion card must be installed starting with the first p5-570 building block.

2.4.2 SP bus

In addition to the processor drawer interconnect cable (described in 2.2.1, “Processor drawer interconnect cables” on page 24), the interconnection of multiple p5-570 building blocks requires the proper SP Flex cable to ensure the vital data communications between the building blocks. (See Figure 2-9 on page 28.) The SP Flex cable contains the system interconnect signals such as JTAG, I2C, clocks, and others.

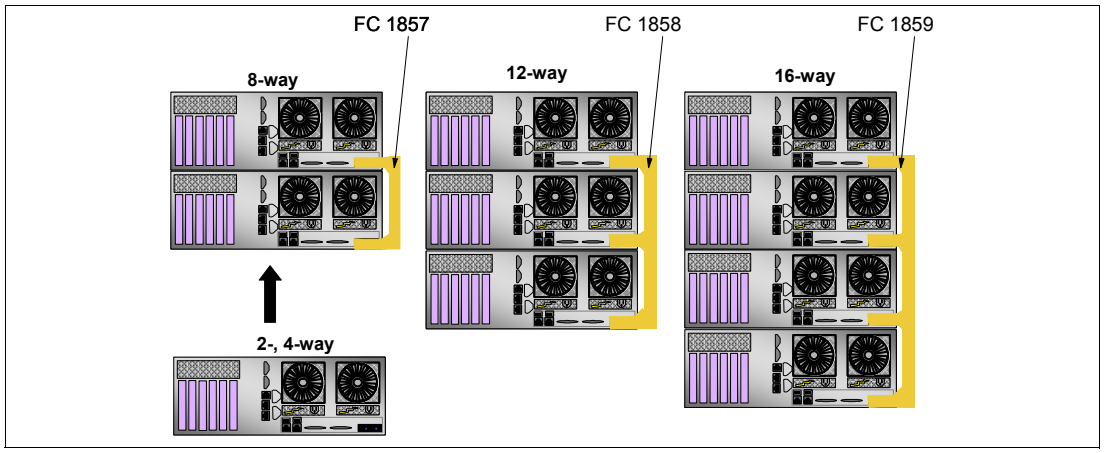


Figure 2-9 FSP Flex cables

2.5 Internal I/O subsystem

The internal I/O subsystem resides on the system planar, and the SP is packaged on a separate service processor card. Each card is a separate FRU. An internal RIO-2 bus is imbedded in the system planar. The system planar contains both the Enterprise RIO-2 hub and the PCI-X Host bridge chip to connect to the integrated I/O that is packaged on the system planar. Two RIO-2 ports of the Enterprise hub chip are used for the integrated I/O and the remaining two ports are routed to external connectors.

2.5.1 PCI-X slots and adapters

PCI-X, where the X stands for extended, is an enhanced PCI bus that delivers a bandwidth of up to 1 GBps, running a 64-bit bus at 133 MHz. PCI-X is backward-compatible, so the p5-570 can support existing 3.3 V PCI adapters.

The system planar provides six PCI-X slots and several integrated PCI devices that interface the three PCI-X to PCI-X bridges to the primary PCI-X buses on the PCI-X Host bridge chip.

PCI-X slot 6 can accept a short PCI-X or PCI card, and its space is shared with the Remote I/O expansion card, therefore if the Remote I/O expansion card is installed, this slot must remain empty. The remaining PCI-X slots are full-length cards. The dual 1 Gb Ethernet adapter is integrated on the system planar.

The PCI-X slots in the p5-570 system support hot-plug and Extended Error Handling (EEH). In the unlikely event of a problem, EEH-enabled adapters respond to a special data packet that is generated from the affected PCI-X slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system reboot.

64-bit and 32-bit adapters

IBM offers 64-bit adapter options for the p5-570 as well as 32-bit adapters. Higher-speed adapters use 64-bit slots because they can transfer 64 bits of data for each data transfer phase. Generally, 32-bit adapters can function in 64-bit PCI-X slots; however, some 64-bit adapters cannot be used in 32-bit slots. For a full list of the adapters that are supported on the p5-570 systems, and for important information regarding adapter placement, visit the IBM @server pSeries Information Center at:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/

2.5.2 LAN adapters

The dual port internal 10/100/1000 Mbps RJ-45 Ethernet controller integrated on the system planar can be used to connect to a local area network (LAN).

Table 2-1 lists additional LAN adapters that were available at the time of writing, when building an initial order of a p5-570 system. IBM supports an installation with NIM using Ethernet and token-ring adapters (CHRP[4] is the platform type.)

Table 2-1 Available LAN adapters

Feature code	Adapter description	Slot	Size
4959	4/16 Token Ring	32 or 64	short
4962	10/100 Ethernet	32 or 64	short
5700	Gigabit Ethernet (Fiber)	64	short
5701	10/100/1000 Ethernet (UTP)	64	short
5706	2-port 10/100/1000 Ethernet (UTP)	64	short
5707	2-port Gigabit Ethernet - SX (Fiber)	64	short
5718	10 Gigabit Ethernet PCI-X	64	short

2.5.3 Graphic accelerators

The p5-570 supports up to two enhanced POWER GXT135P (FC 2849) 2D graphic accelerators. The POWER GXT135P is a low-priced 2D graphics accelerator for pSeries and p5 servers. It can be configured to operate in either 8-bit or 24-bit color modes, running at 60 Hz to 85 Hz. This adapter supports both analog and digital monitors. The adapter requires one short 32-bit or 64-bit PCI-X slot.

2.5.4 SCSI adapters

To connect to external SCSI devices, the adapters listed in Table 2-2 are available, at the time of writing, to be used in p5-570 system.

[4] CHRP stands for Common Hardware Reference Platform, a specification for PowerPC-based systems that can run multiple operating systems.

Table 2-2 Available SCSI adapters

Feature code	Adapter description	Slot	Size
5703	Ultra320 SCSI RAID, bootable	64	long
5712	Ultra320 SCSI	64	short
6204	Ultra SCSI Differential	32	short

2.6 Internal storage

Two Ultra320 SCSI controllers under EADS-X chips that are integrated into the system planar are used to drive the internal disk drives. The six internal drives plug into the disk drive backplane, which has two separate SCSI buses and controllers with three disk drives per bus. Each of these controllers can be dynamically assigned to partitions if required.

The internal disk drive bays can be used in two different modes, depending on whether the SCSI RAID Enablement Card (FC 5709) is installed. (See 2.6.2, “Internal RAID options” on page 31.)

The p5-570 supports a split 6-pack disk drive backplane, which is designed for hot-pluggable disk drives. The disk drive backplane docks directly to the system planar. The virtual SCSI Enclosure Services (VSES) hot plug control functions are provided by the Ultra320 SCSI controllers.

2.6.1 Internal hot swappable SCSI disks

The p5-570 can have up to six hot-swappable disk drives plugged in the two logical 3-pack disk drive backplanes. The hot-swap process is controlled by the virtual SCSI Enclosure Services (VSES), which is located in the logical 3-pack disk drive backplane. (AIX assigns the name `vses0` to the first 3-pack, and `vses1` to the second, if present.) The two logical 3-pack disk drive backplanes can accommodate the devices listed in Table 2-3.

Table 2-3 Hot-swappable disk options

Feature code	Description
3273	36.4 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive
3277	36.4 GB 15,000 RPM Ultra3 SCSI hot-swappable disk drive
3274	73.4 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive
3278	73.4 GB 15,000 RPM Ultra3 SCSI hot-swappable disk drive
3275	146.8 GB 10,000 RPM Ultra3 SCSI hot-swappable disk drive

At the time of writing, if a new order is placed with more than one disk, the system configuration that is shipped from manufacturing may balance the total number of SCSI disks between the two logical 3-pack SCSI backplanes. In this case, this is for manufacturing test purposes and not because of any limitation. Having the disks balanced between the two 3-pack disk drive backplanes enables the manufacturing process to systematically test the SCSI paths and the devices related to them.

Prior to the hot-swap of a disk drive in the hot-swappable-capable bay, all necessary operating system actions must be undertaken to ensure that the disk is capable of being deconfigured. After the disk drive has been deconfigured, the SCSI enclosure device will power-off the slot, enabling safe removal of the disk. You should ensure that the appropriate

planning has been given to any operating-system-related disk layout, such as the AIX Logical Volume Manager, when using disk hot-swap capabilities. For more information, see *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496.

Note: It is recommended that you follow this procedure, after the disk has been deconfigured, when removing a hot-swappable disk drive:

1. Release the tray handle on the disk assembly.
2. Pull out the disk assembly a little bit from the original position.
3. Wait up to 20 seconds until the internal disk stops spinning.

Now you can safely remove the disk from the DASD backplane.

After the SCSI disk hot-swap procedure, you can expect to find `SCSI_ERR10` logged in the AIX error log, with the second word of the sense data equal to `0017`. It is generated from a SCSI bus reset that is issued by the VSES to reset all processes when a drive is inserted, and it is not an issue.

Hot-swap disks and Linux

Linux does not support hot-swap of any disk drive at the time of writing, therefore the Linux operating system does not support these hot-swappable procedures. A p5-570 system running Linux must be shut down and powered off before you replace any disk drives.

2.6.2 Internal RAID options

Every p5-570 building block system is delivered with a disk drive cage that supports up to six disk drive units, offering both internal RAID and non-RAID solutions. When internal RAID solution is not required, at least one 36.4 GB 10K disk drive (FC 3273) is required.

The internal RAID solution requires at least three 36.4 GB 10K disk drives (FC 3273) and the SCSI RAID Enablement Card (FC 5709). Other supported disk drives may be ordered in place of FC 3273. When the SCSI RAID Enablement Card is installed in the system, it re-sequences the two SCSI controllers that support the six disk drive bays, transforming the system from two logical 3-packs of disk drives to one physical 6-pack of disk drives.

The RAID implementation requires a minimum of three disk drives to form a RAID array, so when an order comes in place with FC 5709, at least three disk drives must be in the order list.

Note: Because the p5-570 building block has six disk drive bays, customers performing upgrades must plan accordingly to ensure the correct handling of their RAID arrays.

The p5-570 system supports external RAID solutions, and this requires an additional PCI-X adapter (such as the FC 5703) and external disk drives enclosure.

2.6.3 Internal media devices

The p5-570 provides two slim-line media bays per drawer for optional DVD-ROM (FC 2640) and optional DVD-RAM (FC 5751).

2.7 External I/O subsystems

This section describes the external I/O subsystems, which include the 7311 I/O drawers and the external storage solutions that p5-570 supports.

2.7.1 I/O drawers

As described in Chapter 1, “General description” on page 1, the p5-570 system has six internal PCI-X slots. If more PCI-X slots are needed to dedicate more adapters to a partition or to increase the bandwidth of network adapters, up to 20 7311 model D10, 7311 model D11, and 7311 model D20 I/O drawers can be added to the p5-570 system.

The p5-570 building block system configures a default RIO-2 bus to connect the internal PCI-X slots through the PCI-X to PCI-X bridges, and supports up to four external I/O drawers. To support more I/O drawers in one p5-570 building block, the RIO-2 expansion card (FC 1800) is needed. The RIO-2 expansion card supports up to four additional I/O drawers. If the combined system is made of more than one p5-570 building block, the optional RIO-2 expansion card may be not required if I/O drawers can be shared between the default RIO-2 loop of the p5-570 building blocks.

2.7.2 7311 Model D10 and 7311 Model D11 I/O drawers

The 7311-D10 is not supported in p5-570 initial orders. Only an existing 7311-D10 drawers at the customer site may be migrated from other systems to a p5-570. The 7311-D11 can be in the initial order and it provides six additional PCI-X slots supporting an enhanced blind-swap mechanism. That is the first difference between the two I/O drawer models. The other difference is that the 7311-D11 has six slots that are PCI-X capable, and the D10 has five PCI-X capable slots and one PCI slot. The two types of blind-swap cassette mechanisms are unique for any I/O drawer model; therefore you cannot move a blind-swap cassette from a D10 to a D20 or from a D20 to a D10. Drawers must have a RIO-2 adapter to connect to the server.

Each primary PCI-X bus is connected to a PCI-X-to-PCI-X bridge, which provides three slots with Extended Error Handling (EEH) for error recovering. In the 7311 Model D10 I/O drawer, slot 1 is a standard PCI slot, which operates at 33 MHz and 5 V signaling. In the 7311 Model D11 I/O drawer, slots 1 to 6 are PCI-X slots that operate at 133 MHz and 3.3 V signaling. Figure 2-10 shows a conceptual diagram of the 7311 Model D10 I/O drawer.

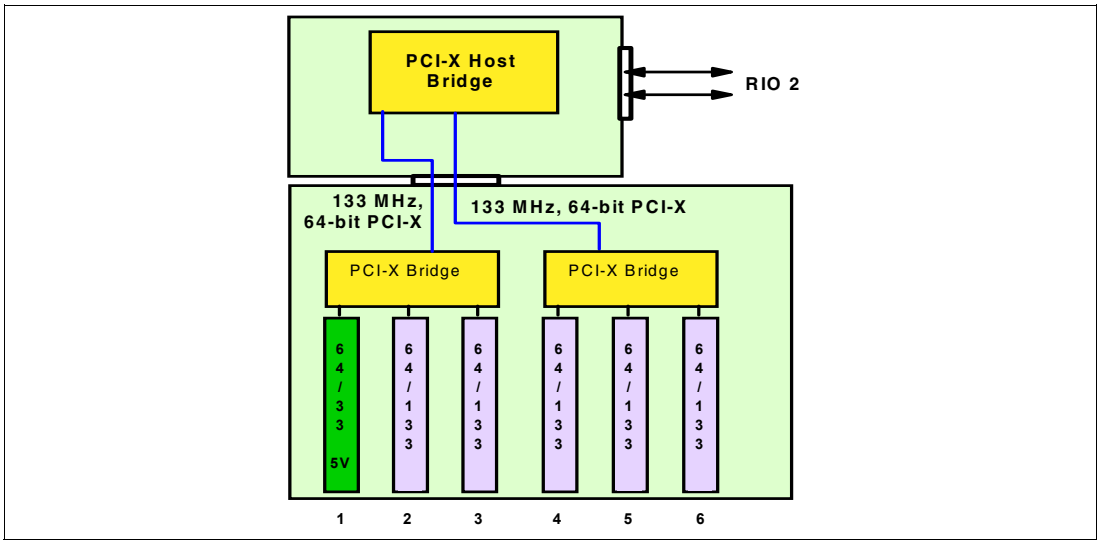


Figure 2-10 Conceptual diagram of the 7311-D10 I/O drawer

7311 Model D10 features

This I/O drawer model provides the following features:

- Six adapters slots:
  - Five hot-plug 64-bit 133 MHz 3.3 V PCI-X slots, full length, blind-swap cassette
  - One hot-plug 64-bit 33 MHz 5 V PCI slot, full length, blind-swap cassette
- Default redundant hot-plug power and cooling
- Two default Remote I/O ports and two SPCN ports

7311 Model D11 features

This I/O drawer model provides the following features:

- Six hot-plug 64-bit, 133 MHz, 3.3 V PCI-X slots, full length, enhanced blind-swap cassette
- Default redundant hot-plug power and cooling
- Two default remote (RIO-2) ports and two SPCN ports

Note: The 7311-D10 must have the RIO- 2 loop adapter (FC 6438) to be supported by a p5-570 system.

2.7.3 7311 Model D20 I/O drawer

The 7311 Model D20 I/O drawer must have the RIO-2 loop adapter (FC 6417) to be connected to the p5-570 system. The PCI-X host bridge inside the I/O drawer provides two primary 64-bit PCI-X buses running at 133 MHz. Therefore, a maximum bandwidth of 1 GBps is provided by each of the buses. To avoid overloading an I/O drawer, follow the recommendation in the IBM @server Hardware Information Center at:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/

Figure 2-11 shows a conceptual diagram of the 7311 Model D20 I/O drawer subsystem.

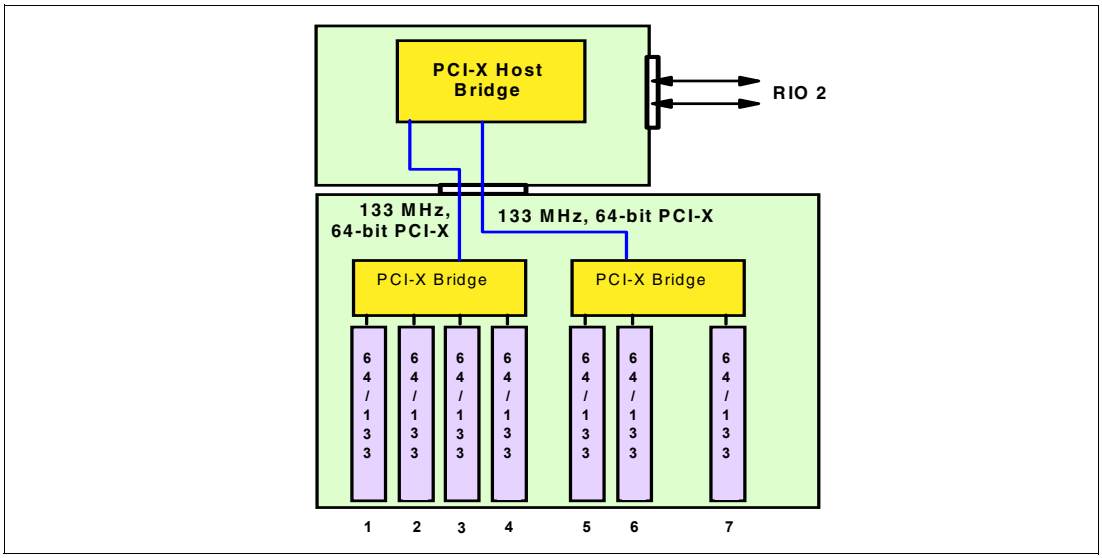


Figure 2-11 Conceptual diagram of the 7311-D20 I/O drawer

7311 Model D20 internal SCSI cabling

A 7311 Model D20 supports hot-swappable disks using two 6-pack disk bays for a total of 12 disks. Additionally, the SCSI cables (FC 4257) are used to connect a SCSI adapter (any of various features) in slot 7 to each of the 6-packs, or two SCSI adapters, one in slot 4 and one in slot 7. (See Figure 2-12 on page 34.)

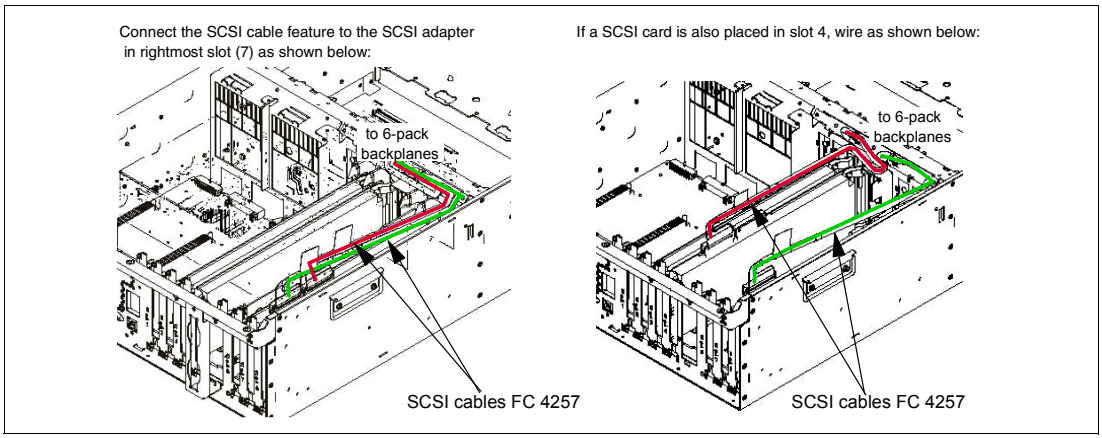


Figure 2-12 7311 Model D20 internal SCSI cabling

Note: Any 6-packs and the related SCSI adapter can be assigned to a partition. If one SCSI adapter is connected to both 6-packs, then both 6-packs can be assigned only to the same partition.

2.7.4 7311 I/O drawer and RIO-2 cabling

As described in 2.7, “External I/O subsystems” on page 32, we can connect up to four I/O drawers in the same loop, and up to 20 I/O drawers to the p5-570 system.

Each RIO-2 port can operate at 1 GHz in bidirectional mode and is capable of passing data in each direction on each cycle of the port. Therefore, the maximum data rate is 4 GBps per I/O drawer in double barrel mode.

There is one default primary RIO-2 loop in any p5-570 building block. This feature provides two Remote I/O ports for attaching up to four 7311 Model D10, 7311 Model D11, or 7311 Model D20 I/O drawers to the system in a single loop. Different I/O drawer models can be used in the same loop, but the combination of I/O drawers must be a total of four per single loop. The optional RIO-2 expansion card may be used to increase the number of I/O drawers that can be connected to one p5-570 building block, and the same rules of the default RIO-2 loop must be considered. The method that is used to connect the drawers to the RIO-2 loop is important for performance. Figure 2-13 on page 35 shows how you could connect four I/O drawers to one p5-570 building block. This is a logical view, actual cables should be wired according to installation instructions.

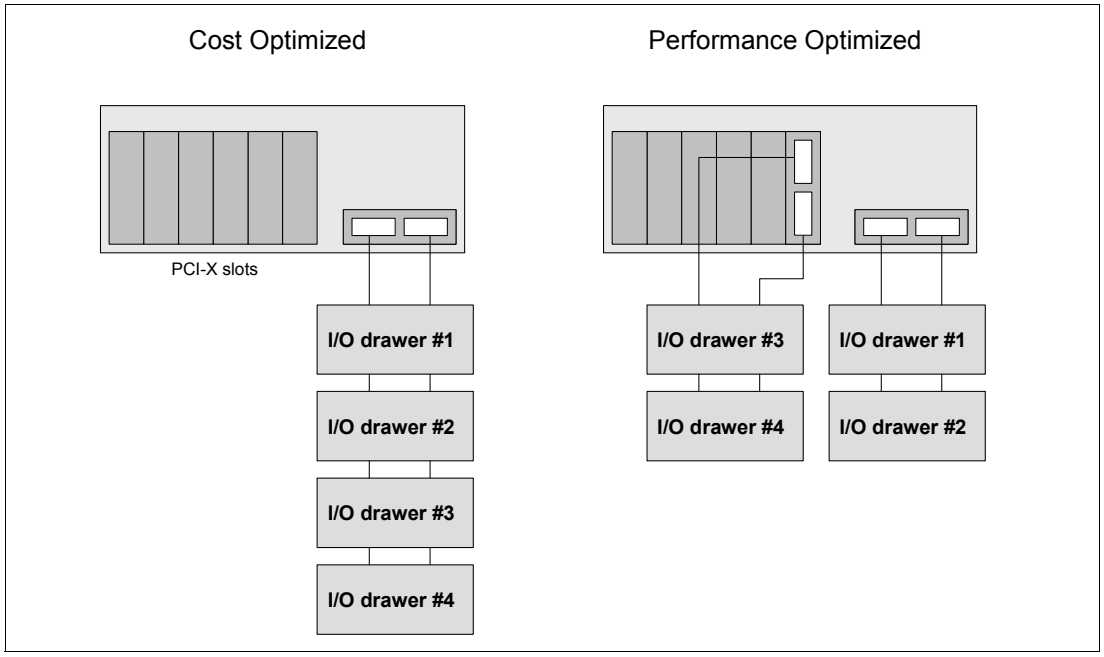


Figure 2-13 RIO-2 cabling examples

Note: If you have 20 I/O drawers, although there are no restrictions on their placement, this can affect performance.

RIO-2 cables are available in different lengths to satisfy different connection requirements:

- Remote I/O cable, 1.2 m (FC 3146, for between D10 and D11 drawers only)
- Remote I/O cable, 3.5 m (FC 3147)
- Remote I/O cable, 10 m (FC 3148)

2.7.5 7311 I/O drawer and SPCN cabling

SPCN[5] is used to control and monitor the status of power and cooling within the I/O drawer. SPCN is a loop: Cabling starts from SPCN port 0 on the p5-570 to SPCN port 0 on the first I/O drawer. The loop is closed, connecting the SPCN port 1 of the I/O drawer back to port 1 of

[5] System Power Control Network

the p5-570 system. If you have more than one I/O drawer, you continue the loop, connecting the next drawer (or drawers) with the same rule.

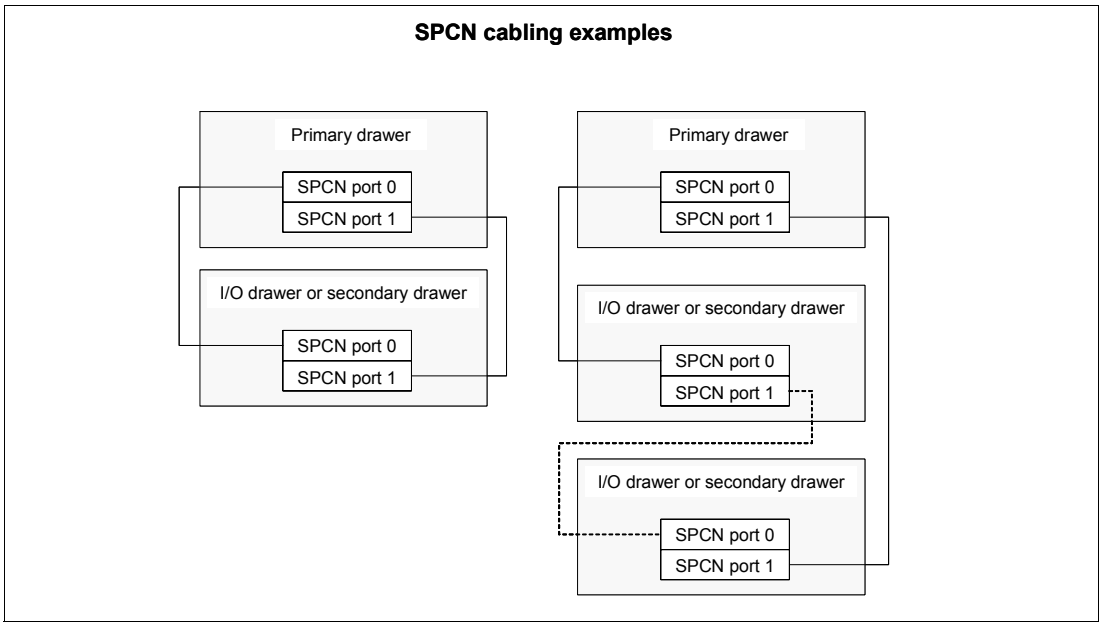


Figure 2-14 SPCN cabling examples

There are different SPCN cables to satisfy different length requirements:

- SPCN cable drawer-to-drawer, 2 m (FC 6001)
- SPCN cable drawer-to-drawer, 3 m (FC 6006)
- SPCN cable rack-to-rack, 6 m (FC 6008)
- SPCN cable rack-to-rack, 15 m (FC 6007)

2.7.6 External disk subsystems

The p5-570 system has internal hot-swappable drives. Internal disks are usually used for the AIX rootvg and paging space. The two 3-pack backplanes can be split when the SCSI RAID Enablement Card is not part of the p5-570 configuration. Additional customer requirements can be satisfied with any of several external disk options that the p5-570 supports.

IBM 2104 Expandable Storage Plus

The IBM 2104 Expandable Storage Plus Model DS4 is a low-cost 3U disk subsystem that supports up to 14 Ultra320 SCSI disks from 18.2 GB to 146.8 GB at time of writing. This subsystem could be used in splitbus mode, which means that the bus with 14 disks could be split into two buses with seven disks each. In this configuration, two additional LPARs could be provided with up to seven disks by using one Ultra3 SCSI adapter (FC 6203) for each LPAR. If RAID is required, one Ultra3 SCSI RAID adapter (FC 5703) can be used for each LPAR.

For more information about the IBM 2104 Expandable Storage Plus subsystem, visit:

http://www.storage.ibm.com/hardsoft/products/expplus/expplus.htm

IBM 7133 Serial Disk Subsystem (SSA)

The p5-570 system supports the Advanced Serial RAID adapter (FC 6230), to attach the IBM 7133 Serial Disk Subsystem Model D40, if migrated from a pSeries system.

The IBM 7133 Serial Disk Subsystem Model D40 provides a 4U highly available storage subsystem for pSeries servers and is a good solution for providing disks for booting additional LPARs. Disks are available from 18.2 GB to 72.8 GB at the time this publication was written. SSA disk subsystems are connected to the server in loops. Each 7133 disk subsystem provides a maximum of four loops with a maximum of four disks each. Therefore, up to four additional LPARs could be provided with disks with dedicated loops for booting by using one Advanced Serial RAID Plus adapter (FC 6230) for each LPAR. Disk space for booting could be provided in JBOD (Just a Bunch Of Disks) or RAID mode.

Notes: FC 6230 serial RAID adapters provide boot support from a RAID-configured disk with firmware level 7000 and above.

Fastwrite cache must not be enabled on the boot resource SSA adapter.

For more information about SSA boot, see the SSA Frequently Asked Questions page at:

`http://www.storage.ibm.com/hardsoft/products/ssa/faq.html`

For further information about SSA, visit the 7133 pages at:

`http://www.storage.ibm.com/hardsoft/products/7133/7133.htm`

IBM TotalStorage FAStT Storage servers

The IBM TotalStorage FAStT Storage server family consists of five models: 100, 200, 600, 700, and 900. As this publication was written Model 100 was the smallest model, scaling to 14 TB. Model 900 is the largest, and it scales to 32 TB. Model 600 provides up to 16 bootable partitions, which are attached with the Gigabit Fibre Channel adapter (FC 6239 or FC 5716). Model 700 provides up to 64 bootable partitions. In most cases, both the FAStT storage server and the p5-570 or the 7311 I/O drawers are connected to a Storage Area Network (SAN). If space is needed only for the rootvg, the FAStT Model 100 or 200 is a good solution.

Note: To boot the p5-570 system, one logical partition, or any other pSeries server from a SAN using FC 6239, the adapter microcode 3.22A1 or later is required. Boot support is provided from direct-attached FastT as well as from SAN attached FastT.

For support of additional features and for further information about the FAStT family, refer to:

`http://www.storage.ibm.com/hardsoft/disk/fastt/index.html`

IBM TotalStorage Enterprise Storage Server®

The IBM TotalStorage Enterprise Storage Server (ESS) is the high-end premier storage solution for use in Storage Area Network. The 2105 Model 800 provides from 582 GB to 55.9 TB of usable disk capacity. An ESS system could also be used to provide disk space for booting LPARs or partitions using Micro-Partitioning technology. An ESS is usually connected to a Storage Area Network (SAN) to which the p5-570 is also connected by using Gigabit Fibre Channel adapters (FC 6239 or FC 5716).

For further information about ESS refer to the following Web site:

`http://www.storage.ibm.com/hardsoft/products/ess/index.html`

2.8 Dynamic logical partitioning

The logical partition (LPAR) was introduced with the POWER4 processor product line and the AIX 5L Version 5.1 operating system. The technology offered the capability to divide a pSeries system into separate systems, where each LPAR runs an operating environment on dedicated attached devices, such as processors, memory, and I/O components. Customers requested system flexibility to change the system topology on demand, and this was achieved by modifying the system layout on the required HMC. Global or individual changes take part on all involved partitions to redefine the new partition layout. Therefore, a reboot of one or more partitions was required.

Later, dynamic LPAR increased the flexibility, enabling selected system resources such as processors, memory, and I/O components to be added and deleted from dedicated partitions while they were executing. Therefore, AIX 5L V5.2 with all necessary enhancements to enable dynamic LPAR was introduced in 2002. This required an attached HMC with the proper level of software to control the system resources, and an updated system firmware level to electronically isolate systems resources. The ability to reconfigure dynamic LPARs encourages system administrators to dynamically redefine all available system resources to reach the optimum capacity for each defined dynamic LPAR.

Dynamic logical partitioning is supported by AIX 5L for POWER V5.2 and later. Its support by SUSE LINUX Enterprise Server 9 and later is planned, but with reduced functionality (changing memory attributes dynamically is not supported, at the time of writing). Dynamic logical partitioning is not supported by the current version of Red Hat Enterprise Linux AS for POWER Version 3.

On the p5-570, the USB devices are considered a group, as are the slimline devices. Devices within a group must be moved from partition to partition as a group. Other devices, such as individual I/O slots, can be relocated individually.

2.9 Virtualization

On the p5-570 server, logical partitions requiring dedicated resources may now be able to take advantage of a new technology that allows resources to be virtualized, allowing for a better overall balance of global system resources and their effective utilization.

2.9.1 Virtual Ethernet

To enhance communication between partitions (dedicated partitions or partitions using Micro-Partitioning technology), the Virtual Ethernet implementation enables in-memory connections at a high bandwidth from partition to partition. Virtual Ethernet working on LAN technology enables a transmission speed in the range of 1 GBps to 3 GBps, depending on the MTU[6] size, and it supports 256 Virtual Ethernet connections in a partition, where a single Virtual Ethernet resource can be connected to another Virtual Ethernet, a real network adapter, or both in a partition.

2.9.2 Advanced POWER Virtualization feature

The Advanced POWER Virtualization feature is an optional additional cost hardware feature that is available on all IBM ℮server POWER5 processor-based systems. Each system has a unique feature code for this feature. For the p5-570 server, select FC 7942 to order the Advanced Virtualization feature.

[6] Maximum Transmission Unit

The Advanced POWER Virtualization feature includes:

- Firmware enablement for Micro-Partitions
- Installation image for the Virtual I/O server software that supports:
  - Ethernet adapter sharing
  - Virtual SCSI Server
- Partition Load Manager
  - Automated CPU and memory reconfiguration
  - Real-time partition configuration and load statistics
  - Graphical user interface

Micro-Partitioning technology

POWER5-based servers introduces an enhanced partitioning model that is based on the partitioning concepts of a stable and well-known mainframe technology and on existing LPAR/dynamic LPAR implementation on POWER4 and POWER4+ servers.

The Micro-Partitioning model offers a virtualization of system resources. In POWER5 processor-based systems, physical resources are abstracted into virtual resources that are available to partitions. This sharing method is the primary feature of this new partitioning concept and it happens transparently.

POWER5 Micro-Partitioning specifies processor capacity in processing units. One processing unit represents 1% of one physical processor. 1.0 represents the power of one processor. A partition defined with 220 processing units is equivalent to the power of 2.2 physical processors. Creating a partition using Micro-Partitioning technology, the minimum capacity is 10 processing units, or 1/10 of a physical processor. A maximum of 10 partitions for each physical processor may be defined, but on a loaded system the practical limit is less. In a p5-570 system with 16 processors in a shared pool, up to 160 partitions using Micro-Partitioning technology can be activated at the same time for an entire system. The practical limit to the number of partitions is based on available hardware and performance objectives.

Micro-Partitions can also be defined with capped and uncapped attributes. A capped Micro-Partition is not allowed to exceed the defined capacity (a configuration flag inside the HMC menus determines whether the capacity is capped), while an uncapped partition is allowed to consume additional capacity with fewer restrictions. Uncapped partitions can be configured to the total idle capacity of the server or a percentage of it.

The POWER5 processor-based systems use the POWER Hypervisor, which is the new Hypervisor for executing the Micro-Partition model. The Hypervisor of existing POWER4 processor-based systems works on a demand basis, as the result of machine interrupts and callbacks to the operating system. The new Hypervisor operates continuously in the background.

The Advanced POWER Virtualization Feature, which is described in 2.9.2, “Advanced POWER Virtualization feature” on page 38, facilitates all POWER5 and POWER Hypervisor enhancements to reach the highest level of granularity of installed system resources.

Figure 2-15 shows the POWER5 partitioning concept.

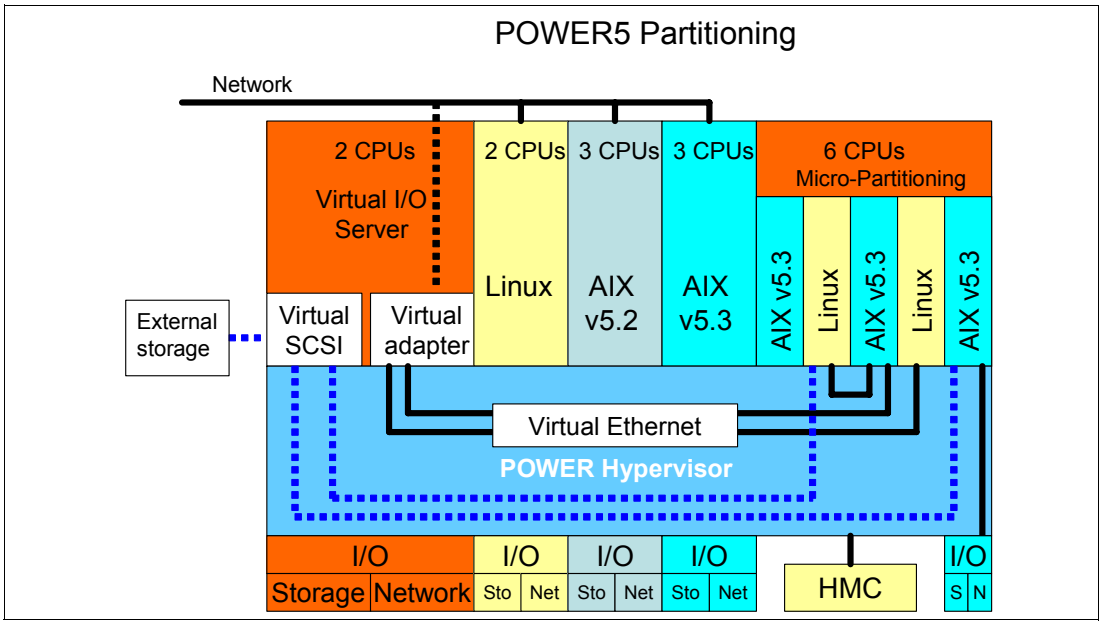


Figure 2-15 Micro-Partitioning and LPAR

Virtual I/O Server

The Virtual I/O Server is a special-purpose partition that provides virtual I/O resources to client partitions. The Virtual I/O Server owns the real resources that are shared with the other clients. With Virtual I/O technology, a physical adapter can be assigned to a partition to be shared by one or more partitions, enabling clients to minimize their number of physical adapters. The Virtual I/O Server can be used to reduce costs by eliminating the requirement that each partition has a dedicated network adapter, disk adapter, and disk drive.

It is preferable to use the Virtual I/O server in a partition with dedicated resources to help ensure stable performance.

The following sections discuss the two major functions that are provided from the Virtual I/O Server.

Shared Ethernet adapter

A Shared Ethernet Adapter is a new service that acts as a Layer 2 network switch to route network traffic from a Virtual Ethernet to a real network adapter. The Shared Ethernet Adapter must run in a Virtual I/O Server partition.

The advantage of the Virtual Ethernet Services is that partitions to communicate outside the system without having a physical network adapter attached to the partition. Up to 18 VLANs can be shared on a single network interface. The amount of network traffic will limit the number of client partitions that are served through a single network interface.

Virtual SCSI

Access to real storage devices is implemented through the Virtual SCSI services, a part of the Virtual I/O Server partition. Logical volumes that are created and exported on the Virtual I/O Server partition will be shown at the virtual storage client partition as a SCSI disk. All current storage device types such as SAN, SCSI, and RAID are supported.

The Virtual I/O server supports logical mirroring, and RAID configurations. Logical volumes created on RAID or JBOD configurations are bootable, and the number of logical volumes is limited to the amount of storage available and architectural limits of the LVM.

Note: The Shared Ethernet adapter and Virtual SCSI server functions are provided in the Virtual I/O Server that is included in the Advanced POWER Virtualization feature (FC 7942), an additional feature of p5 systems.

Partition Load Manager

The Partition Load Manager (PLM) is part of the Advanced POWER Virtualization feature. It provides automated processor and memory distribution between dynamic LPAR and Micro-Partitioning capable logical partition running AIX 5L. The PLM application is based on a client/server model to share system information, such as processor or memory events, across the concurrent present logical partitions.

To improve the overall resource utilization of a partitioned system, PLM uses user-defined resource management policies to determine the additional resources, such as processors and memory, for each requesting partition.

For network communication, the PLM uses the Resource Monitoring and Control (RMC) subsystem, which provides several events on every managed node. The following events are registered on all managed nodes:

- Memory-pages-steal high thresholds and low thresholds
- Memory-usage-high thresholds and low thresholds
- Processor-load-average high threshold and low threshold

To ensure a secure communication between managed nodes, OpenSSH and Kerberos V5 are supported in PLM to have a secure communication and an authentication mechanism for administrators. If Kerberos is not installed, PLM uses the configured authentication method, such as AIX authentication.

2.10 Service processor

The service processor (SP) is an embedded controller that is based on a PowerPC 405GP processor (PPC405) running the SP internal operating system. The SP operating system contains specific programs and device drivers for the SP hardware.

The p5-570 includes the SP. The key components include an FSP-Base (FSP-B) and an Extender chipset (FSP-E). FSP-B and FSP-E are implemented on a dedicated card.

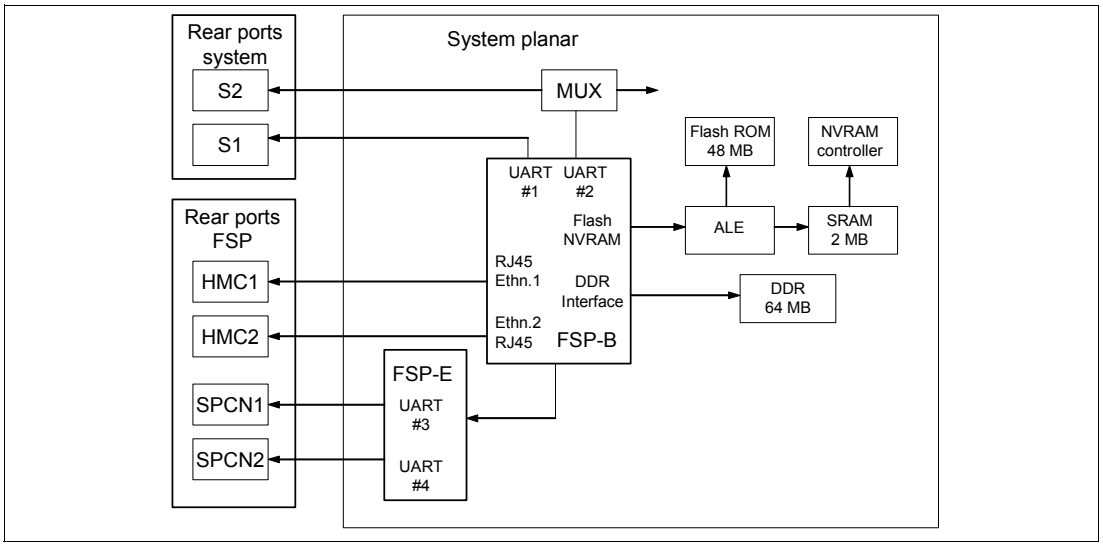


Figure 2-16 Service processor logic diagram

2.10.1 Service processor - base

The PPC405 core features a five-stage pipeline instruction processor and contains 32-bit general purpose registers. The flash ROM contains a compressed image of a software load.

The SP base unit offers the following connections:

- Two Ethernet Media Access Controller3 (MAC3) cores, which is a generic implementation of the Ethernet Media Access (MAC) protocol compliant with ANSI/IEEE 802.3, IEEE 802.3u, ISO/IEC 8802.3 CSMA/CD Standard. The Ethernet MAC3 supports both half-duplex (CSMA/CD) and full-duplex operation at 10/100 Mbps. Both Ethernet ports are only visible to the service processor.
- Two serial interfaces, which are accessible only though the serial ports of p5-570 on the rear side. At the time of writing, the System Management Interface (SMI) is usable if a connection is established to serial port 1. Terminals that are connected to serial port 2 receive only boot sequence information without manual interaction. When the HMC is connected to the SP, the serial ports are disabled and do not provide any external connection.

2.10.2 Service processor - extender

The SP extender unit offers the two system power control network (SPCN) ports to control the power of the attached I/O subsystem. The SPCN control software and the service processor software are run on the same PPC405 processor.

2.11 Boot process

From the earlier RS/6000 systems, through the previous pSeries systems, the boot process passed through several enhancements. With the implementation of the POWER5 technology, the boot process is enhanced to accommodate the flexibility that the POWER5 processor-based hardware features. Depending on the customer's needs, a system may or may not require the use of an HMC to manage the system. The boot process, based on the Initial Program Load (IPL) setup, is determined by the hardware setup and the way you use the features that POWER5 processor-based systems provide.

The IPL process starts when power is connected to the system. Immediately after, the SP starts an internal self test (Built-In-Self-Test, or BIST) that is based on integrated diagnostic programs. The system status changes to standby only when all of the test units have passed.

2.11.1 IPL flow without an HMC attached to the system

When system status is standby, the SP presents a System Management Interface (SMI), which can be accessed by striking any key on an attached serial console keyboard, or the Advanced System Management Interface (ASMI), which uses a Web browser[7] on a client system that is connected to the SP on an Ethernet network.

The SP and the ASMI are standard on all POWER5 processor-based hardware. Both system management interfaces require the general or admin ID password, and they both enable you to set flags that affect the operation of the system according to the provided password, such as auto-power restart, view information about the system (such as the error log and VPD), network environment access setup, and control of system power.

You can start and shut down the system in addition to setting IPL options. The p5-570 has a permanent firmware boot side, or A side, and a temporary firmware boot side, or B side. New levels of firmware should be installed on the temporary side first in order to test the update's compatibility with your applications. When the new level of firmware has been approved, it can be copied to the permanent side.

In the SMI and ASMI, you can view and change IPL settings:

- System boot speed
  - Fast or Slow: Fast boot results in skipped diagnostic tests and shorter memory tests during the boot.
- Firmware boot side for next boot
  - Permanent or Temporary: Firmware updates should be tested by booting from the temporary side before being copied into the permanent side.
- System operating mode
  - Manual or Normal: Manual mode overrides various automatic power-on functions, such as auto-power restart, and enables the power switch button.
- AIX/Linux partition-mode boot (available only if the system is not managed by the HMC)
  - Service mode boot from saved list: This is the preferred way to run concurrent AIX diagnostics.
  - Service mode boot from default list: This is the preferred way to run stand-alone AIX diagnostics.

[7] Supported browsers are Netscape (version 7.1), Internet Explorer (version 6.0), and Opera (version 7.23). At the time of writing, older or previous versions of these browsers are not supported. JavaScript™ and cookies must be enabled.

- Boot to open firmware prompt
- Boot to System Management Service (SMS) to further select the boot devices or network boot options

- Boot to server firmware
  - Select the state for the server firmware: Standby or Running.
  - When the server is in the server firmware standby state, partitions can be set up and activated.

2.11.2 Hardware Management Console

Depending on the model, the HMC provides several native serial ports and Ethernet ports. One serial port may be used to attach a modem for Service Agent. Service Agent Connection Manager can be used instead, if the HMC has a TCP/IP port 80 connection to the Internet. The HMC provides an Ethernet port (or ports) to connect to partitions on its managed POWER5 processor-based systems. The network connection is mandatory for the HMC to p5 systems, and highly recommended between the HMC and partitions. It supports the following functions:

- Logical partition configuration and management
- Dynamic Logical Partitioning
- Capacity and resource management
- System status
- HMC management
- Service functions (for example, microcode updates and Service Focal Point)
- Remote HMC interface

Note: The same HMC cannot be attached to POWER4 and POWER5 processor-based systems simultaneously, but for redundancy purposes one POWER5 server can be attached to two HMCs.

All managed servers must be authenticated from the HMC. If a new attached system is discovered, the HMC will prompt you to set two passwords using the HMC interface:

- Advanced System Management general user ID password
- Advanced System Management admin ID password
- HMC access password

2.11.3 IPL flow with an HMC attached to the system

When system status is standby, you can either use the HMC to open a virtual terminal and access the SMI, or launch a Web browser to access the ASMI.

Using the SMI or the ASMI, you can view or modify the proper IPL settings in order to set the boot mode to partition standby and then turn the system on, but the HMC can be also used to power on the managed system (and is highly recommended). Using the HMC to turn the system on requires selecting one of the following:

- Partition Standby

  The Partition Standby power-on mode enables you to create and activate logical partitions.

- When the Partition Standby power-on is completed, the operator panel on the managed system displays *LPAR...*, indicating that the managed system is ready for you to use the HMC to partition its resources and, possibly, activate them.
- When a partition is activated, the HMC requires you to select the boot mode of the single partition.

- System Profile

  The System Profile option powers on the system according to a predefined set of profiles. Profiles are activated in the order in which they are shown in the system profile.

- Partition autostart

  This option powers on the managed system to partition standby mode and then activates all partitions that have been designated autostart.

After the system succeeds to boot with any of these choices, the HMC can be used to manage the system, such as continuing to boot from operating system or managing the logical partitions. See 2.11.2, “Hardware Management Console” on page 44.

2.11.4 Definitions of partitions

Describing the detailed process for working with the HMC and the management tasks to create and manage a logical partition, LPAR, or dynamic LPAR is not the intention of this documentation. The following section describes the additional functionality to create partitions that use fractional elements of available system resources, Micro-Partitioning.

For better understanding of the partitioning concept, this section gives an overview of common terminology. In addition to the partitioning concept, there are two components:

- Managed Systems
- Profiles

Managed systems

Managed systems are physical systems that are managed by the HMC. One HMC can manage more managed systems at a time.

Profiles

A profile defines the configuration of a logical partition or managed system. There are three types of profiles that can be used to create multiple profiles for each logical partition or managed system:

- Partition profile
  - A partition profile includes the collection of resource specifications, such as processing units, memory, and I/O resources, because a logical partition is not aware of a resource until it is activated.
  - A logical partition can have more than one partition profile, but at least one is a minimum requirement.
- All resources dedicated partition profile

  A partition profile that contains the entire resource list of the machine, using all physical resources working as one system.

- System profile
  - A system profile is an ordered list of partition profiles. When you activate a system profile, the managed system attempts to activate the partition profiles in the defined order.

– To enhance the flexibility to use the system within several different logical configurations, a System profile could be defined to collect more than one Partition profile to provide requested system behavior.

2.11.5 Hardware requirements for partitioning

To implement partitioning on a POWER5-based system, resource planning is important to ensure that you have a base configuration and enough flexibility to make desirable changes to the running logical partitions. To configure a partition, the minimum requirements are processors, memory, and virtual or physical resources to provide boot, application and network support.

Processors

Within POWER5 technology and depending on performance requirements, a logical partition can be created by using a shared processor pool or a dedicated processor.

Shared processors can be defined by a fractional number starting at 1/10th the capacity of a real processor. To calculate the required processor power, a real processor is divided in 100 processing units, so 1/10 of a processor is equal to 10 processing units.

Dedicated processors are entire processors that can be assigned to a single logical partition without the capability to share free capacity to other logical partitions.

Memory

Depending on given application and performance requirements, a logical partition requests memory to execute the installed operating system and application. To create partitions, the minimum memory requirement is 128 MB per logical partition and dynamically increases in increments of 16 MB from the overall memory available in the system.

Expansion unit

Expansion units extend the flexibility of the server system to increase the number of possible logical partitions by adding additional hardware, such as storage or network devices.

2.11.6 Specific partition definitions used for Micro-Partitioning

In addition to the base definition for a partition, new parameters must be defined in order to achieve more flexibility of partitions using Micro-Partitioning technology.

Capped and uncapped partition

A capped partition indicates that the local partition will never exceed its assigned capacity. An uncapped partition indicates that if the capacity entitlement is reached, additional capacity from the shared pool can be used if available.

To manipulate the behavior of uncapped partitions, the parameter uncapped weight must be defined, in the range from 0 through 255. To prevent an uncapped partition from receiving extra capacity, the uncapped weight parameter should be 0.

The default uncapped weight is 128.

2.11.7 System Management Services

Booting up a full partition system or a logical partition to System Management Services (SMS) using the ASCII[8] interface or the GUI results in identical contents and functionality.

The p5-570 (or the logical partition) must be equipped with either a graphic adapter that is connected to a graphics display, keyboard, and mouse device, or an ASCII display terminal that is connected to one of the native serial ports or the attached HMC to use the SMS menus. You can view information about the system (or the single logical partition) and perform tasks such as set a password, change the boot list, and set the network parameters.

If the system or the partition has been activated without flagging the option to stop to the SMS, you have the option to press the 1 key on the terminal or in the graphic window after the word `keyboard` appears and before the word `speaker` appears. In the terminal or in the GUI, the system or the partitions require you to enter the password that is defined for admin or general access. When the text-based SMS starts (either for terminal or graphic window), a screen similar to Figure 2-17 opens.

```
Version SF220_004
SMS 1.5 (c) Copyright IBM Corp. 2000,2003 All right reserved
------------------------------------------------------------------------------
Main Menu
 1.  Select Language
 2.  Setup Remote IPL (Initial Program Load)
 3.  Change SCSI Settings
 4.  Select Console
 5.  Select Boot Options









------------------------------------------------------------------------------
Navigation Keys:

                                          X = eXit System Management Services
------------------------------------------------------------------------------
Type the number of the menu item and press Enter or select Navigation Key:
```

Figure 2-17 System Management Services main menu

Note: The version of system firmware that is installed in your system is displayed at the top of each screen. Processor and other device upgrades might require a specific version of firmware to be installed in your system.

On each menu screen, you are given the option of choosing a menu item and pressing Enter (if applicable) or selecting a navigation key. You can use the different options to review or set the boot list information, or to set up the network environment parameters if you want the system to boot from a NIM server.

2.11.8 Boot options

The p5-570 handles the boot process in a way that is similar to other pSeries servers.

The initial stage of the boot process establishes that the machine has powered up correctly and that the memory and CPUs are functioning correctly. When the machine or the logical

[8] American Standard Code for Information Interchange: this is the worldwide standard for the code numbers that are used by computers to represent all uppercase and lowercase Latin letters, numbers, punctuation, and so on.

partition displays the SMS menus, all of the necessary tests have been performed and the machine is scanning the bus for a boot source.

Most system backplanes are designed such that the drive in the first slot spins up immediately after power-on, and other drives will wait for the operating system to send a command before spinning up. Disk drive bays 1 and 4 are hard-wired to spin up immediately. The left-most slot of the 3-pack disk backplanes (SCSI ID 5, boot, autostart) is set to spin up immediately after power-on. The power-on delay sequence is performed to prevent power supply overloading. This behavior makes the disk in the first slot of the first 3-pack DASD backplane the preferred boot device. See Figure 2-18 to locate all of the disk bays.

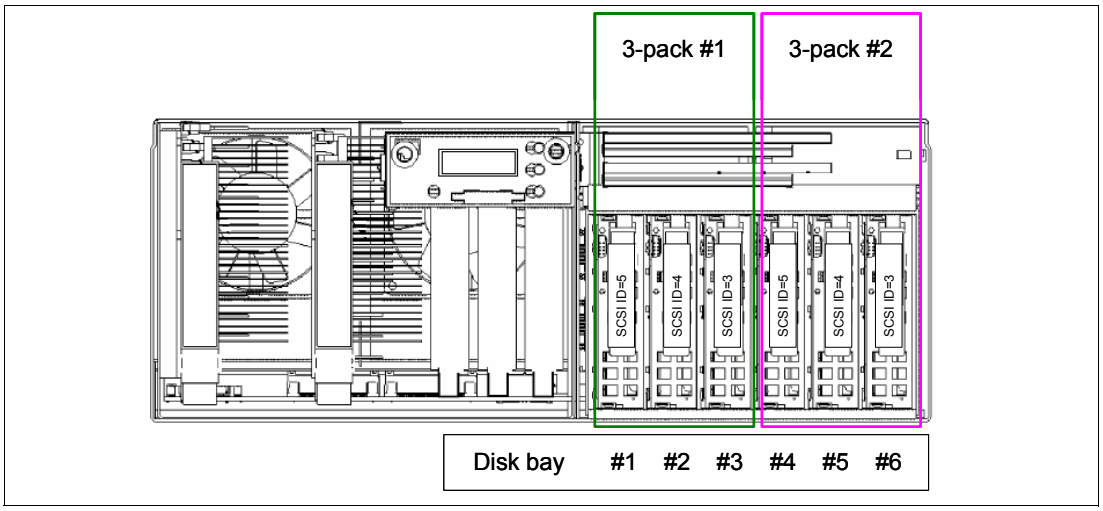


Figure 2-18 Disk bays and SCSI addresses within a p5-570 building block

When SMS menus are available, the Select Boot Options menu can be used to view and set various options regarding the installation devices and boot devices:

1. Select Install or Boot a Device

   Enables you to select a device to boot from or install the operating system from. This selection is only for the current boot.

2. Select Boot Devices

   Enables you to set the boot list.

3. Multiboot Startup

   Toggles the multiboot startup flag, which controls whether the multiboot menu is invoked automatically on startup.

2.11.9 Additional boot options

Instead of booting from the preferred boot device or from any other internal disks, there are a number of other possibilities:

DVD-ROM, DVD-RAM These devices can be used to boot the system or a logical partition (if the resource is available to the specific partition), so that a system can be loaded, or system maintenance or stand-alone diagnostics can be performed.

External tape drives Any externally attached tape drive can be used to boot the system, or a logical partition if the resource is

available to the specific partition, using **mksysb**, for example.

SCSI disk, and Virtual SCSI disk The more common method of booting the system is to use a disk situated in one of the hot-swap bays in the front of the machine. However, any external SCSI-attached disk could be used if required. As described in previous sections, Virtual SCSI devices are also available to a logical partition.

SSA disk The p5-570 supports booting from an SSA disk either as an AIX system disk or as a RAID LUN. For more information about the SSA boot, see the SSA Frequently Asked Questions located on the Web[9].

Important: Fastwrite must not be enabled on the boot resource SSA adapter.

SAN boot It is possible to boot the p5-570 system from an SAN using a 2 GB Fibre Channel Adapter (FC 6239 or FC 5716), or to boot one partition using the dedicated 2 GB Fibre Channel Adapter or the Virtual SCSI device related to this adapter. The IBM 2105 Enterprise Storage Server (ESS) is an example of an SAN-attached device that can provide a boot medium.

LAN boot Network boot and NIM installs can be used if required. Logical partitions can use either a dedicated Ethernet adapter or Virtual Ethernet to accomplish this.

2.11.10 Security

The p5-570 system enables you to set two different types of passwords to limit access to these systems: The *admin* password is usually used by the system administrator, and the *general ID password* provides limited access to the system functions and is usually available to all users who are allowed to power-on the server, especially remotely. These are defined in the ASMI menus.

2.12 Operating system requirements

All new POWER5 servers are capable of running IBM AIX 5L for POWER and support appropriate versions of Linux. AIX 5L has been specifically developed and enhanced to exploit and support the extensive RAS features on IBM @server pSeries systems.

2.12.1 AIX 5L

The p5-570 requires AIX 5L Version 5.3 or AIX 5L Version 5.2 Maintenance Package 5200-04 (IY56722) or later.

The system requires the following media:

- AIX 5L for POWER Version 5.2 5765-E62, dated 08/2004 (CD# LCD4-1133-04) or later
- AIX 5L for POWER Version 5.3 5765-G03, dated 08/2004 (CD# LCD4-7463-00) or later

[9] http://www.storage.ibm.com/hardsoft/products/ssa/faq.html

IBM periodically releases maintenance packages for the AIX 5L operating system. These packages are available on CD-ROM (FC 0907) and can be downloaded from the Internet at:

http://techsupport.services.ibm.com/server/fixes

You can also get individual operating system fixes and information about obtaining AIX 5L service at this site. AIX5L Version 5.3 has the **suma** command, which helps the administrator to automate the task of checking and downloading operating system downloads.

If you have problems downloading the latest maintenance level, ask your IBM Business Partner or IBM representative for assistance.

The Advanced POWER Virtualization feature is not supported on AIX 5L Version 5.2.

2.12.2 Linux

For the p5-570, Linux distributions were available through SUSE and Red Hat at the time this publication was written. The p5-570 requires the following version of Linux distributions:

- SUSE LINUX Enterprise Server 9 for POWER systems, or later
- Red Hat Enterprise Linux AS for POWER Version 3

The Advanced POWER Virtualization feature, DLPAR, and other features require SUSE SLES 9. Red Hat Enterprise Linux supports the Advanced POWER Virtualization feature.

In Japan, Turbolinux is also available. In the Latin America sales region, Conectiva is also available. For related information and an overview, see:

http://www.ibm.com/servers/eserver/pseries/linux

Find full information about Red Hat Enterprise Linux AS for POWER Version 3 at:

http://www.redhat.com/software/rhel/as/

Find full information about SUSE Linux Enterprise Server 9 for POWER at:

http://www.suse.com/us/business/products/server/sles/i_pseries.html

For information about UnitedLinux for pSeries from Turbolinux, see:

http://www.turbolinux.co.jp

For the latest in IBM Linux news, subscribe to the Linux Line. See:

https://www6.software.ibm.com/reg/linux/linuxline-i

Many of the features that are described in this document are OS-dependent and may not be available on Linux. For more information, check:

http://www.ibm.com/servers/eserver/pseries/linux/whitepapers/linux_pseries.html

Linux support

IBM only supports the Linux systems of customers with a SupportLine contract that covers Linux. Otherwise, the Linux distributor should be contacted for support.

3

Capacity on Demand, RAS, and manageability

The following sections provide more detailed information about IBM @server p5 design features that help lower total cost of ownership (TCO). This chapter includes several features that are based on the benefits that are available when using AIX 5L. Support of these features using Linux can vary.

3.1 Capacity on Demand

p5-570 systems can be shipped with non-activated resources (processors and memory), which may be added as they are needed. Processors and memory can be brought online to meet increasing workload demands without affecting system operations.

The CoD is supported on p5 Model 570 when the 1.65 GHz and 1.9 GHz Power5 processors are used.

The following sections outline the different methods that are available, namely:

- Processor
  - Capacity Upgrade on Demand
  - Reserve Capacity on Demand
  - Dynamic processor sparing
- Memory
  - Permanent Capacity Upgrade on Demand for Memory
  - On/Off Capacity on Demand
  - Reserve Capacity on Demand
- Processor and Memory
  - Trial Capacity on Demand for processors and memory

The following sections describe the processor cards that support the CUoD features.

2-way 1.65 GHz POWER5 processor card

The base 2-way 1.65 GHz POWER5 processor card (FC 7830) features two processors (but 0 active) and eight DDR1 Memory DIMM sockets. Processors are activated in increments of one processor, either permanently or for a given amount of time.

2-way 1.9 GHz POWER5 processor card with DDR1 memory slots

The 2-way 1.9 GHz POWER5 processor card (FC 7832) features two processors (but 0 active) and eight DDR1 Memory DIMM sockets. Processors are activated in increments of one processor, either permanently or for a given amount of time.

2-way 1.9 GHz POWER5 processor card with DDR2 memory slots

The 2-way 1.9 GHz POWER5 processor card (FC 7833) features two processors (but 0 active) and eight DDR2 Memory DIMM sockets. Processors are activated in increments of one processor, either permanently or for a given amount of time.

Note: The 2-way 1.5 POWER5 processor card does not support any CUoD feature.

Processor and memory Capacity Upgrade On Demand and Dynamic processor sparing is supported by AIX 5L Version 5.2 ML4 (IY56722) or AIX 5L Version 5.3. Dynamic processor sparing requires that the CPU guard attribute is set to enable.

3.1.1 Processor Capacity Upgrade on Demand methods

This section describes the different CUoD methods that are available for the processors at time of writing.

Capacity Upgrade on Demand

In Capacity Upgrade on Demand (CUoD), processors are shipped to the customer as installed in the p5-570, and can be activated later in increments of one processor. Additional options deliver the possibility to temporarily use the processor resources that are installed in the server.

All processor cards are 2-way, with 0-way active. In an initial order of the p5-570 system, at least two processors must be activated by ordering the appropriate activation features. At least 2 GB of installed memory must be installed on each processor card independent of processor activation.

On/Off Capacity on Demand

After an On/Off Processor Enablement feature is ordered and the associated enablement code is entered into the system, customers must report On/Off usage to IBM at least monthly. This information, which is used to compute the billing data, is then provided to the sales channel, which will place an order for a quantity of On/Off Processor Day Billing features.

Reserve Capacity on Demand

The Reserve Capacity option activation is a reserve capacity for 30 processor-days of prepaid reserve. To establish processor capacity on the server, select a quantity of inactive processors to be placed in the server's shared processor pool as reserve processors. After the server recognizes that non-reserve (permanently activated) processors that are assigned or available to the uncapped partitions have been 100% utilized, use of additional processors will cause processor days (good for a 24-hour period) to be subtracted from the prepaid number of processor days.

The 30-days Reserve Capacity Processor feature is activated with FC 7956.

Dynamic Processor Sparing

When you have non-activated CUoD processors, a feature called Dynamic Processor Sparing is automatically provided. Non-activated CUoD processors are processors that are physically installed in the system but not activated. Dynamic Processor Sparing makes the system capable of disabling a failing processor and enabling a non-activated CUoD available processor.

3.1.2 Capacity Upgrade on Demand for memory

Capacity Upgrade on Demand for memory offers the ability to add memory resources and to activate in increments of 1 GB using encrypted keys.

Each of these features consists of a set of four DIMMs, with half of the memory already activated and half of the memory available for later activation. Memory can be activated in different ways using feature codes described in “Capacity Upgrade on Demand for memory feature codes” on page 54.

Permanent activation

CUoD for memory is activated in increments of 1 GB and can be activated permanently by ordering FC 7950. The FC 7950 provides 1024 MB activation for DDR1 memory.

CUoD memory features are supported only on 1.65 GHz and 1.9 GHz processor cards with DDR1 memory.

On/Off Capacity on Demand

The On/Off memory enablement feature is ordered and the associated enablement code is entered into the system. On/off usage must be reported to IBM at least monthly. This information, which is used to compute your billing data, is then provided to the sales channel, which will place an order for a quantity of On/Off Memory day billing features. One FC 7957 should be ordered for each billable day and for each 1GB increment of DDR1 memory.

On/Off memory enablement can be activated with FC 7954.

Capacity Upgrade on Demand for memory feature codes

This section describes the CUoD for memory activation feature codes and provides some examples of memory granularity depending on the activation feature codes.

The feature codes to activate the memory are:

- Base memory CUoD, 8 GB (4 GB activated): FC 7890
- On/Off enablement memory: FC 7954
- 1 GB memory activation: FC 7950
- CUoD 1 GB memory: FC 7952
- On/Off one day billing: FC 7957

Based on these feature codes, use Table 3-1 to see examples of increments and granularity that CUoD for memory offers, compared to base installed memory (non-CUoD feature codes).

Table 3-1 Granularity of memory with CUoD feature code compared to standard memory

Memory required (GB)	Base installed memory FCs	CUoD FCs	One-Day Activation FCs
4	4453 (DDR1), or 7893 (DDR2)	7890	7890 7954
6	not applicable	7890 and 7950 x 2	7890 7954 7957 x 2
7	not applicable	7890 7950 x 3	7890 7954 7957 x 3
8	(FC 4454 or FC 7894)	7890 7950 x 4	7890 7954 7957 x 4

3.1.3 How to report temporary activation resources

There are three methods for reporting information about usage of On/Off Capacity on Demand to IBM (contact your areas IBM representative for the numbers and addresses):

- Using Electronic Service Agent™

  Monthly reporting of temporary capacity billing information can be sent to IBM electronically using the Electronic Service Agent, which is part of the Hardware Management Console, and is designed to monitor events and to transmit server inventory information to IBM on a periodic, customer-definable timetable.

- Using fax

- Using e-mail

3.1.4 Trial Capacity on Demand

Customers with CUoD featured systems must purchase the activation codes from IBM before the non-activated CUoD resources can be activated to meet the increased workload. However with the Trial Capacity on Demand feature, customers can activate the required non-activated CUoD resources immediately and, after that, proceed to purchase those resources from IBM or not. The HMC calls this feature *Activate Immediate*. A one-time no-cost activation for a maximum period of 30 consecutive days is available as a complementary service when access to CUoD resources is required immediately.

The following basic rules apply for the p5-570 system:

- After the CUoD resources are activated, the customer must either buy all or part of the activated CUoD resources from IBM or return the activated CUoD resources to the system within 30 days.
- Trial CUoD can only be used once.

There are several advantages to using this feature:

- You can improve the response time to meet unpredictable increase in workload.
- Customer can monitor the performance of the system after activating the CUoD resources before placing the order for activation codes.
- It is useful when a CUoD permanent activation purchase is pending.

3.2 Reliability, availability, and serviceability

Excellent quality and reliability are inherent in all aspects of the IBM @server p5 design and manufacturing. The fundamental objective of the design approach is to minimize outages. The RAS features help to ensure that the system operates when required, performs reliably, and efficiently handles any failures that may occur. This is achieved by using capabilities that are provided by both the hardware and AIX 5L.

The p5-570 as a POWER5 server improves on the RAS capabilities that are implemented in POWER4-based systems. Available RAS enhancements on POWER5 servers are:

- Most firmware updates enable the system to remain operational.
- ECC has been extended to inter-chip connections for the Fabric and Processor bus.
- Partial L2 cache deallocation is possible.
- Number of L3 cache line deletes improved from 2 to 10 for better self-healing capability.

The following sections describe the concepts that form the basis of leadership RAS features of IBM @server p5 systems in more detail.

3.2.1 Fault avoidance

p5 systems are built to keep errors from ever happening. This quality-based design includes such features as:

- Reduced power consumption and cooler operating temperatures for increased reliability, enabled by the use of copper chip circuitry, SOI, and dynamic clock-gating
- Mainframe-inspired components and technologies

3.2.2 First Failure Data Capture

If a problem should occur, the ability to diagnose it correctly is a fundamental requirement upon which improved availability is based. The p5-570 incorporates advanced capability in start-up diagnostics and in run-time First Failure Data Capture (FDDC) based on strategic error checkers built into the chips.

Any errors that are detected by the pervasive error checkers are captured into Fault Isolation Registers (FIRs, shown in Figure 3-1), which can be interrogated by the service processor (SP). The SP in the p5-570 has the capability to access system components using special-purpose service processor ports or by access to the error registers.

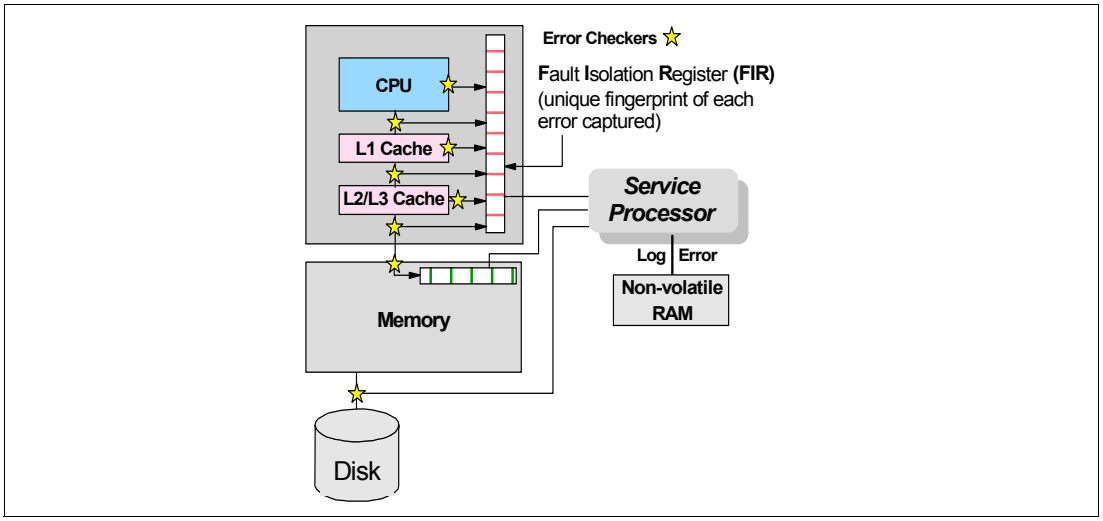


Figure 3-1 Schematic of Fault Isolation Register implementation

The FIRs are important because they enable an error to be uniquely identified, thus enabling the appropriate action to be taken. Appropriate actions might include such things as a bus retry, ECC correction, or system firmware recovery routines. Recovery routines could include dynamic deallocation of potentially failing components.

Errors are logged into the system non-volatile random access memory (NVRAM) and the SP event history log, along with a notification of the event to AIX for capture in the operating system error log. Diagnostic Error Log Analysis (*diagela*) routines analyze the error log entries and invoke a suitable action such as issuing a warning message. If the error can be recovered, or after suitable maintenance, the service processor resets the FIRs so that they can accurately record any future errors.

The ability to correctly diagnose any pending or firm errors is a key requirement before any dynamic or persistent component deallocation or any other reconfiguration can take place.

3.2.3 Permanent monitoring

The SP that is included in the p5-570 provides a way to monitor the system even when the main processor is inoperable. The next subsection offers a more detailed description of monitoring functions in p5-570.

Mutual surveillance

The SP can monitor the operation of the firmware during the boot process, and it can monitor the operating system for loss of control. This enables the service processor to take appropriate action, including calling for service, when it detects that the firmware or the

operating system has lost control. Mutual surveillance also enables the operating system to monitor for service processor activity and can request a service processor repair action if necessary.

Environmental monitoring

Environmental monitoring related to power, fans, and temperature is performed by the System Power Control Network (SPCN). Environmental critical and non-critical conditions generate Early Power-Off Warning (EPOW) events. Critical events (for example, a Class 5 AC power loss) trigger appropriate signals from hardware to affected components so as to prevent any data loss without operating-system or firmware involvement. Non-critical environmental events are logged and reported using Event Scan.

The operating system cannot program or access the temperature threshold using the SP.

EPOW events can trigger the following actions:

- Temperature monitoring, which increases the fan's speed rotation when ambient temperature is above a preset operating range.
- Temperature monitoring warns the system administrator of potential environmental-related problems. It also performs an orderly system shutdown when the operating temperature exceeds a critical level.
- Voltage monitoring provides warning and an orderly system shutdown when the voltage is out of operational specification.

3.2.4 Self-healing

For a system to be self-healing, it must be able to recover from a failing component by first detecting and isolating the failed component, taking it offline, fixing or isolating it, and reintroducing the fixed or replacement component into service without any application disruption. Examples include:

- *Bit steering* to redundant memory in the event of a failed memory module to keep the server operational
- *Bit-scattering*, thus allowing for error correction and continued operation in the presence of a complete chip failure (*Chipkill™ recovery*)
- Single-bit error correction using ECC without reaching error thresholds for main, L2, and L3 cache memory
- L3 cache line deletes extended from 2 to 10 for additional self-healing
- ECC extended to inter-chip connections on fabric and processor bus
- *Memory scrubbing* to help prevent soft-error memory faults
- *Dynamic processor deallocation,* in which a deallocated processor can be replaced by an unused CoD processor to keep the system operational

Memory reliability, fault tolerance, and integrity

The p5-570 uses Error Checking and Correcting (ECC) circuitry for system memory to correct single-bit memory failures and to detect double-bit. Detection of double-bit memory failures helps maintain data integrity. Furthermore, the memory chips are organized such that the failure of any specific memory module only affects a single bit within a four-bit ECC word (*bit-scattering*), thus allowing for error correction and continued operation in the presence of a complete chip failure (*Chipkill recovery*). The memory DIMMs also utilize *memory scrubbing* and thresholding to determine when spare memory modules within each bank of memory should be used to replace ones that have exceeded their threshold of error count

(*dynamic bit-steering*). Memory scrubbing is the process of reading the contents of the memory during idle time and checking and correcting any single-bit errors that have accumulated by passing the data through the ECC logic. This function is a hardware function on the memory controller chip and does not influence normal system memory performance.

3.2.5 N+1 redundancy

The use of redundant parts allows the p5-570 to remain operational with full resources:

- Redundant spare memory bits in L1, L2, L3, and main memory
- Redundant fans
- Redundant power supplies

3.2.6 Fault masking

If corrections and retries succeed and do not exceed threshold limits, the system remains operational with full resources and no client or IBM customer engineer intervention is required:

- CEC bus retry and recovery
- PCI-X bus recovery
- ECC Chipkill soft error

3.2.7 Resource deallocation

If recoverable errors exceed threshold limits, resources can be deallocated with the system remaining operational, allowing deferred maintenance at a convenient time.

Dynamic or persistent deallocation

Dynamic deallocation of potentially failing components is non-disruptive, allowing the system to continue to run. Persistent deallocation occurs when a failed component is detected, which is then deactivated at a subsequent reboot.

Dynamic deallocation functions include:

- Processor
- L3 cache line delete
- Partial L2 cache deallocation
- PCI-X bus and slots

For dynamic processor deallocation, the service processor performs a predictive failure analysis based on any recoverable processor errors that have been recorded. If these transient errors exceed a defined threshold, the event is logged and the processor is deallocated from the system while the operating system continues to run. This feature (named *CPU Guard*) enables maintenance to be deferred until a suitable time. Processor deallocation can occur only if there are sufficient functional processors (at least two).

To verify whether CPU Guard has been enabled, run the following command:

```
lsattr -El sys0 | grep cpuguard
```

If CPU Guard is enabled, the output will be similar to:

```
cpuguard    enable      CPU Guard     True
```

If the output shows CPU Guard as disabled, enter the following command to enable it:

```
chdev -l sys0 -a cpuguard='enable'
```

Cache or cache-line deallocation is aimed at performing dynamic reconfiguration to bypass potentially failing components. This capability is provided for both L2 and L3 caches. Dynamic run-time deconfiguration is provided if a threshold of L1 or L2 recovered errors is exceeded.

In case of an L3 cache run-time array single-bit solid error, the spare chip resources are used to perform a L3 cache line delete on the failing line.

PCI hot-plug slot fault tracking helps prevent slot errors from causing a system machine check interrupt and subsequent reboot. This provides superior fault isolation, and the error affects only the single adapter. Run-time errors on the PCI bus that are caused by failing adapters will result in recovery action. If this is unsuccessful, the PCI device will be gracefully shut down. Parity errors on the PCI bus itself will result in bus retry and, if uncorrected, the bus and any I/O adapters or devices on that bus will be deconfigured.

The p5-570 supports PCI Extended Error Handling (EEH) if it is supported by the PCI-X adapter. In the past, PCI bus parity errors caused a global machine check interrupt, which eventually required a system reboot in order to continue. In the p5-570 system, hardware, system firmware, and AIX interaction has been designed to allow transparent recovery of intermittent PCI bus parity errors and graceful transition to the I/O device available state in the case of a permanent parity error in the PCI bus.

EEH-enabled adapters respond to a special data packet that is generated from the affected PCI slot hardware by calling system firmware, which examines the affected bus, allows the device driver to reset it, and continues without a system reboot.

Persistent deallocation functions include:

- Processor
- Memory
- Deconfigure or bypass failing I/O adapters
- L3 cache

Following a hardware error that has been flagged by the service processor, the subsequent reboot of the system invokes extended diagnostics. If a processor or L3 cache has been marked for deconfiguration by persistent processor deallocation, the boot process will attempt to proceed to completion with the faulty device automatically deconfigured. Failing I/O adapters will be deconfigured or bypassed during the boot process.

Note: The auto-restart (reboot) option, when enabled, can reboot the system automatically following an unrecoverable software error, software hang, hardware failure, or environmentally induced failure (such as loss of power supply).

3.2.8 Serviceability

By increasing service productivity, the system is up and running for a longer time. p5-570 improves service productivity by providing the following functions.

Error indication and LED indicators

The p5-570 is designed to be installed by an IBM service representative. The addition of most hardware features after the install is customer setup. To help the customer and the IBM service representative, the p5-570 provides internal LED diagnostics that identify parts that require service. Indication of an error is provided through a series of light attention signals,

starting on the exterior of the system (System Attention LED) and ending with an LED near the failing Field Replaceable Unit.

For more information about replaceable units, including videos, see:

http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/new.htm#cru

System Attention LED

The attention indicator is represented externally by an amber LED on the operator panel and the back of the system unit. It is used to indicate that the system is in one of the following states:

- Normal state: LED is off.
- Fault state: LED is on solid.
- Identify state: LED is blinking.

Additional LEDs on I/O components, such as PCI-X slots and disk drives, provide status information such as power, hot-swap, and need for service.

Concurrent Maintenance

Concurrent Maintenance provides replacement of the following parts while the system remains running:

- Disk drives
- Cooling fans
- Power Subsystems
- PCI-X adapter cards

3.3 Manageability

Functions and tools that are provided for IBM @server p5 systems are described in the following sections.

3.3.1 Advanced System Management Interface

With the system in power standby mode, or with an operating system in control of the machine or controlling the related partition, the SP is working and checking the system for errors, ensuring the connection to the HMC for manageability purposes.

With the system up and running, the SP provides the ability to view and change the power-on settings using the Advanced System Management Interface (ASMI). Also, the surveillance function of the SP is monitoring the operating system to confirm that it is still running and has not stalled.

Figure 3-2 on page 61 shows an example of the ASMI accessed from the Web browser.

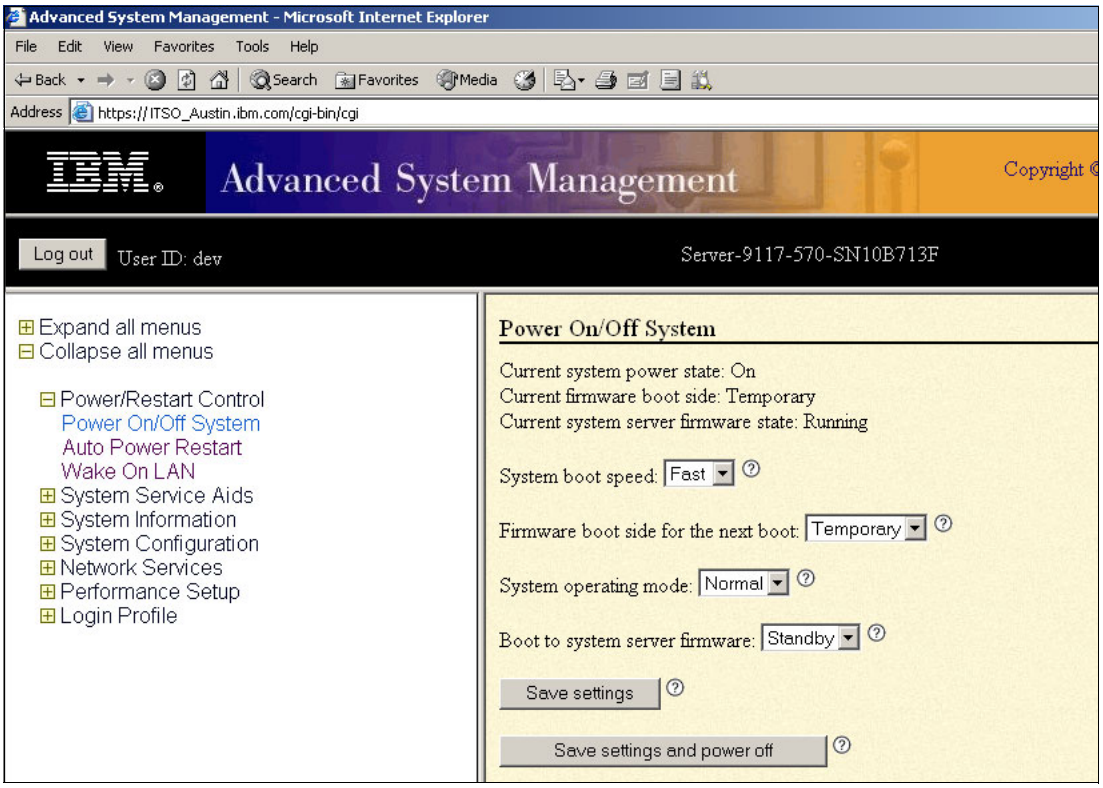


Figure 3-2 Advanced System Management main menu

3.3.2 Service Agent

Service Agent is an application program that operates on a p5, pSeries, or IBM RS/6000 computer and monitors them for hardware errors. It reports detected errors, assuming that they meet certain criteria for criticality, to IBM for service with no customer intervention. It is an enhanced version of Service Director™ with a graphical user interface.

Key things you can accomplish by using Service Agent for p5, pSeries, and RS/6000 are:

- Automatic problem analysis
- Problem-definable threshold levels for error reporting
- Automatic problem reporting; service calls placed to IBM without intervention
- Automatic customer notification

In addition to these features, the following are available:

- Commonly viewed hardware errors: You can view hardware event logs for any monitored machine on the network from any Service Agent host user interface.
- High-availability cluster multiprocessing (HACMP) support for full fallback: includes high-availability cluster workstation (HACWS) for 9076.
- Network environment support with minimum telephone lines for modems.
- VPD data can be sent to IBM using Performance Management.

Machines are defined by using the Service Agent user interface. After machines are defined, they are registered with the IBM Service Agent Server (SAS). During the registration process, an electronic key is created that becomes part of your resident Service Agent program. This key is used each time Service Agent places a call for service. The IBM Service Agent Server

checks the current customer service status from the IBM entitlement database; if this reveals that you are not under warranty or MA, then the service call is refused and posted back using e-mail.

Service focal point

Service Focal Point is used by service technicians to start and end their service calls. It provides service representatives with event, Vital Product Data (VPD), and diagnostic information. The HMC can also notify service representatives of hardware failures automatically by using the Service Agent features. You can configure the HMC to use the Service Agent call-home feature to send IBM event information. This information is stored, analyzed, and then acted on by the service representative. Some parts of Service Focal Point have to be configured so that the proper information is sent to IBM.

You can download the latest version of Service Agent at:

`ftp://ftp.software.ibm.com/aix/service_agent_code`

3.3.3 p5 Customer-Managed Microcode

The pSeries and RS/6000 Customer-Managed Microcode is a methodology that enables you to manage and install microcode updates on p5, pSeries, and RS/6000 systems and associated I/O adapters. The IBM pSeries Microcode Update Web site can be found at:

`http://techsupport.services.ibm.com/server/mdownload`

IBM provides service tools that can assist in determining microcode levels and updating systems with the latest available microcode. To determine which tool to use in a specific environment, visit:

`http://techsupport.services.ibm.com/server/mdownload/mcodetools.html`

3.3.4 Service Update Management Assistant

The Service Update Management Assistant (SUMA) helps system administrators retrieve maintenance updates from the Web. SUMA offers flexible options that enable customers to set up policies to automate the download of fixes to their systems. SUMA policies can be scheduled to periodically check the availability of specific new fixes (APAR, PTF, or fileset), critical or security fixes, or an entire maintenance level. A notification e-mail can be sent detailing updates that are needed when comparing available fixes to installed software, a fix repository, or a maintenance level.

Benefits provided by SUMA include:

- Moves administrators away from the task of manually retrieving maintenance updates from the Web
- Policy can be scheduled to run periodically; for example, download the latest critical fixes weekly
- Can compare fixes needed against software inventory, fix repository, or a maintenance level
- Receive mail notification after a fileset preview or download operation
- Allows for FTP, HTTP, or secure HTTPS transfers
- Provides same requisite checking as the IBM fix distribution Web site
- Available through SMIT menus (**`smitty suma`**) or a command line interface

3.4 Cluster 1600

Today's IT infrastructure requires that systems meet increasing demands while offering the flexibility and manageability to rapidly develop and deploy new services. IBM clustering hardware and software provide the building blocks, with availability, scalability, security, and single-point-of-management control, to satisfy these needs.

IBM *e*server Cluster 1600 is a POWER-based AIX 5L and Linux Cluster targeting scientific and technical computing, large-scale databases, and workload consolidation.

IBM Cluster Systems Management (CSM) is designed to provide a robust, powerful, and centralized way to manage a large number of POWER5-based systems from a single point of control. CSM can help lower the overall cost of IT ownership by helping to simplify the tasks of installing, operating, and maintaining clusters of servers. CSM can provide one consistent interface for managing both AIX and Linux nodes (physical systems or logical partitions), with capabilities for remote parallel network install, remote hardware control, and distributed command execution.

The p5-570 is supported with the Cluster 1600 running CSM for AIX, V1.3.1. To attach a p5-570 to a Cluster 1600, an HMC is required. One HMC can also control several p5-570s that are part of the cluster. If a p5-570 that is configured in partition mode (with physical or virtual resources) is part of the cluster, all partitions must be part of the cluster.

It is not possible to use selected partitions as part of the cluster and use others for non-cluster use. The HMC uses a dedicated connection to the p5-570 to provide the functions that are needed to control the server, such as powering the system on and off. The HMC must have an Ethernet connection to the Control Work Station (CWS). Each partition in p5-570 must have an Ethernet adapter to connect to the CWS *trusted* LAN.

Information about HMC control, cluster building block servers, and available cluster software can be found in the following link:

http://www.ibm.com/servers/eserver/clusters/

The benefits of clustered environment based on logical partitions

Evolving processor and storage technologies has had a great impact on the architecture of IT infrastructures. This was the most significant challenge for the infrastructure in the past and will continue to be in the future. During the first half of the 1990s, one central instance of an application per node was suitable; moreover, most productive systems needed additionally associated nodes, so called application servers.

Increasing performance and reliability by simply replicating application server nodes led to complex environments that often resulted in poor system management. The reason for these complex constructions was the limited computing power of a single node. This limitation was softened during the second half of the 1990s.

Big symmetric multiprocessor (SMP) nodes with higher clock rates and increased memory provided the ability to install more than one system on a node. This had some side effects regarding systems operations: a release's planning processes had to pay attention to different databases, application versions, or both to avoid unresolved conflicts.

In 2000, Workload Manager for AIX (WLM) was announced. Multiple application instance installations became more and more popular because of the permanently increasing number of systems that were dedicated to applications at customer sites. The general availability of this functionality of AIX to separate the workloads of dedicated systems eliminated the last obstacle for consolidating several systems into one node.

Some customers expanded the use of their dedicated systems and consequently model more business processes. This often caused an increased number of dedicated systems that were used and a stronger demand on flexibility. In addition, the life cycle of these systems differed extremely. Renaming, removal, and deletion became more and more common system administration tasks.

In 2001, the pSeries hardware technology with logical partitioning was generally available. Logical partitioning creates the ability to define the logical partitions (LPARs) that are adapted to customer needs regarding the number of processors, assigned memory, and I/O adapters: no waste of resources, but the flexibility to assign the right power at the right moment. The p5-570 offers the flexibility to increase the usage of the resources even more, and reduce the total cost of ownership (TCO).

Partitions with associated physical resources or virtual resources are not different from a collection of stand-alone nodes.

Today, server consolidation is a must for many IT sites. Minimized TCO and complexity, with the maximum amount of flexibility, is a crucial goal of nearly all customers. LPARs enable flexible distribution of resources with LPAR boundaries. Each logical partition can be configured according to the specific needs of the occupant application. LPARs provide a protection boundary between the systems. More test and development systems can exist on the same server in separate partitions.

CSM value points

The CSM allows the management of different hardware platforms from a single point of control, and it has consistent interfaces to manage systems and logical partitions that are running both AIX and Linux. Management is achieved across multiple switch and interconnect topologies. PSSP forced system administrators to do some things a certain way (such as NIM, and SP user management). The CSM provides assistance in setting these things up, but enables system administrators to tailor their systems to their own needs, and it has the ability to manage systems across different geographical sites.

Monitoring is much easier to use, and the system administrator can monitor all of the network interfaces, not just the switch and administrative interfaces. The management server pushes information out to the nodes, which releases the management server from having to trust the node. In addition, the nodes do not have to be network-connected to each other. This means that giving root access on one node does not mean giving root access on all nodes. The base security setup is all done automatically at install time.

The CSM ships with AIX itself (a 60-day Try and Buy license is shipped with AIX). The CSM client side is automatically installed and ready when you install AIX, so each system or logical partition is cluster-ready.

CSM V1.4 on AIX and Linux (planned 4Q04)

The CSM V1.4 on AIX and Linux introduces an optional IBM CSM High Availability Management Server (HA MS) feature, which is designed to allow automated failover of the CSM management server to a backup management server. In addition, sample scripts for setting up NTP[1], and network tuning (AIX ONLY) configurations, and the capability to copy files across nodes or node groups in the cluster can improve cluster ease of use and site customization.

[1] Network Time Protocol

Related publications

The publications listed in this section are considered particularly suitable for a more detailed discussion of the topics covered in this Redpaper.

IBM Redbooks

For information about ordering these publications, see “How to get IBM Redbooks” on page 67. Note that some of the documents referenced here may be available in softcopy only.

- *IBM @server pSeries 670 and pSeries 690 System Handbook*, SG24-7040
- *The Complete Partitioning Guide for IBM pSeries Servers*, SG24-7039
- *Managing AIX Server Farms*, SG24-6606
- *Practical Guide for SAN with pSeries*, SG24-6050
- *Problem Solving and Troubleshooting in AIX 5L*, SG24-5496
- *Understanding IBM @server pSeries Performance and Sizing*, SG24-4810

Other publications

These publications are also relevant as further information sources:

- *7014 Series Model T00 and T42 Rack Installation and Service Guide*, SA38-0577, contains information regarding the 7014 Model T00 and T42 Rack, in which this server can be installed.
- *7316-TF3 17-Inch Flat Panel Rack-Mounted Monitor and Keyboard Installation and Maintenance Guide*, SA38-0643, contains information regarding the 7316-TF3 Flat Panel Display, which can be installed in your rack to manage your system units.
- *IBM @server Hardware Management Console for pSeries Installation and Operations Guide*, SA38-0590, provides information to operators and system administrators on how to use a IBM Hardware Management Console for pSeries (HMC) to manage a system. It also discusses the issues associated with logical partitioning planning and implementation.
- *Planning for Partitioned-System Operations*, SA38-0626, provides information to planners, system administrators, and operators about how to plan for installing and using a partitioned server. It also discusses some issues associated with the planning and implementing of partitioning.
- *RS/6000 and @server pSeries Adapters, Devices, and Cable Information for Multiple Bus Systems*, SA38-0516, contains information about adapters, devices, and cables for your system. This manual is intended to supplement the service information found in the *Diagnostic Information for Multiple Bus Systems* documentation.
- *RS/6000 and @server pSeries Diagnostics Information for Multiple Bus Systems*, SA38-0509, contains diagnostic information, service request numbers (SRNs), and failing function codes (FFCs).
- *RS/6000 and pSeries PCI Adapter Placement Reference*, SA38-0538, contains information regarding slot restrictions for adapters that can be used in this system.
- *System Unit Safety Information*, SA23-2652, contains translations of safety information used throughout the system documentation.

Online resources

These Web sites and URLs are also relevant as further information sources:

- AIX 5L operating system maintenance packages downloads

  http://www.ibm.com/servers/eserver/support/pseries/aixfixes.html
- Autonomic computing on IBM @server pSeries servers

  http://www.ibm.com/autonomic/index.shtml
- Ceramic Column Grid Array (CCGA), see IBM Chip Packaging

  http://www.ibm.com/chips/micronews
- Copper circuitry

  http://www.ibm.com/chips/technology/technologies/copper/
- Frequently asked SSA-related questions

  http://www.storage.ibm.com/hardsoft/products/ssa/faq.html
- Hardware documentation

  http://publib16.boulder.ibm.com/pseries/en_US/infocenter/base/
- IBM @server Information Center

  http://publib.boulder.ibm.com/eserver/
- IBM @server pSeries and RS/6000 microcode update

  http://techsupport.services.ibm.com/server/mdownload2/download.html
- IBM @server pSeries support

  http://www.ibm.com/servers/eserver/support/pseries/index.html
- IBM @server support: Tips for AIX administrators

  http://techsupport.services.ibm.com/server/aix.srchBroker
- IBM Linux news: Subscribe to the Linux Line

  https://www6.software.ibm.com/reg/linux/linuxline-i
- Information about UnitedLinux for pSeries from Turbolinux

  http://www.turbolinux.co.jp
- IBM online sales manual

  http://www.ibmlink.ibm.com
- Linux for IBM @server pSeries

  http://www.ibm.com/servers/eserver/pseries/linux/
- Microcode Discovery Service

  http://techsupport.services.ibm.com/server/aix.invscoutMDS
- POWER4 system micro architecture, comprehensively described in the *IBM Journal of Research and Development*, Vol 46 No.1 January 2002

  http://www.research.ibm.com/journal/rd46-1.html
- SCSI T10 Technical Committee

  http://www.t10.org
- Silicon-on-insulator (SOI) technology

  http://www.ibm.com/chips/technology/technologies/soi/

- SSA boot FAQ

  http://www.storage.ibm.com/hardsoft/products/ssa/faq.html#microcode
- SUSE LINUX Enterprise Server 8 for pSeries information

  http://www.suse.de/us/business/products/server/sles/i_pseries.html
- The LVT is a PC based tool intended assist you in logical partitioning

  http://www-1.ibm.com/servers/eserver/iseries/lpar/systemdesign.htm

How to get IBM Redbooks

You can search for, view, or download Redbooks, Redpapers, Hints and Tips, draft publications and Additional materials, as well as order hardcopy Redbooks or CD-ROMs, at this Web site:

ibm.com/redbooks

Help from IBM

IBM Support and downloads

ibm.com/support

IBM Global Services

ibm.com/services

IBM @server p5 570 Technical Overview and Introduction

Redpaper

Finer system granulation using Micro-Partitioning technology to help lower TCO

Modular midrange solution for managing on demand business

Extreme flexibility with outstanding performance

This document is a comprehensive guide covering the IBM @server p5 570 UNIX servers. We introduce major hardware offerings and discuss their prominent functions.

Professionals wishing to acquire a better understanding of IBM @server p5 products should consider reading this document. The intended audience includes:

- Customers
- Sales and marketing professionals
- Technical support professionals
- IBM Business Partners
- Independent software vendors

This document expands the current set of IBM @server documentation by providing a desktop reference that offers a detailed technical description of the p5-570 system.
This publication does not replace the latest pSeries marketing materials and tools. It is intended as an additional source of information that, together with existing sources, can be used to enhance your knowledge of IBM server solutions.

INTERNATIONAL TECHNICAL SUPPORT ORGANIZATION

BUILDING TECHNICAL INFORMATION BASED ON PRACTICAL EXPERIENCE

IBM Redbooks are developed by the IBM International Technical Support Organization. Experts from IBM, Customers and Partners from around the world create timely technical information based on realistic scenarios. Specific recommendations are provided to help you implement IT solutions more effectively in your environment.

For more information:
ibm.com/redbooks